Hierarchical Object-Based Visual Attention for Machine Vision

Yaoru Sun

Doctor of Philosophy

Institute of Perception, Action and Behaviour
School of Informatics
University of Edinburgh

2003

Abstract

Human vision uses mechanisms of covert attention to selectively process interesting information and overt eye movements to extend this selectivity ability. Thus, visual tasks can be effectively dealt with by limited processing resources. Modelling visual attention for machine vision systems is not only critical but also challenging. In the machine vision literature there have been many conventional attention models developed but they are all space-based only and cannot perform object-based selection. In consequence, they fail to work in real-world visual environments due to the intrinsic limitations of the space-based attention theory upon which these models are built. The aim of the work presented in this thesis is to provide a novel human-like visual selection framework based on the object-based attention theory recently being developed in psychophysics. The proposed solution – a Hierarchical Object-based Attention Framework (HOAF) based on grouping competition, consists of two closely-coupled visual selection models of (1) hierarchical object-based visual (covert) attention and (2) object-based attention-driven (overt) saccadic eye movements. The Hierarchical Object-based Attention Model (HOAM) is the primary selection mechanism and the Object-based Attention-Driven Saccading model (OADS) has a supporting role, both of which are combined in the integrated visual selection framework HOAF.

This thesis first describes the proposed object-based attention model HOAM which is the primary component of the selection framework HOAF. The model is based on recent psychophysical results on object-based visual attention and adopted grouping-based competition to integrate object-based and space-based attention together so as to achieve object-based hierarchical selectivity. The behaviour of the model is demonstrated on a number of synthetic images simulating psychophysical experiments and real-world natural scenes. The experimental results showed that the performance of our object-based attention model HOAM concurs with the main findings in the psychophysical literature on object-based and space-based visual attention. Moreover, HOAM has outstanding hierarchical selectivity from far to near and from coarse to fine by features, objects, spatial regions, and their groupings in complex natural scenes. This successful performance arises from three original mechanisms in the model: grouping-based saliency evaluation, integrated competition between groupings, and hierarchical selectivity. The model is the first implemented machine vision model of integrated object-based and space-based visual attention.

The thesis then addresses another proposed model of Object-based Attention-Driven Saccadic eye movements (OADS) built upon the object-based attention model HOAM,

as an overt saccading component within the object-based selection framework HOAF. This model, like our object-based attention model HOAM, is also the first implemented machine vision saccading model which makes a clear distinction between (covert) visual attention and overt saccading movements in a two-level selection system – an important feature of human vision but not yet explored in conventional machine vision saccading systems. In the saccading model OADS, a log-polar retina-like sensor is employed to simulate the human-like foveation imaging for space variant sensing. Through a novel mechanism for *attention-driven orienting*, the sensor fixates on new destinations determined by object-based attention. Hence it helps attention to selectively process interesting objects located at the periphery of the whole field of view to accomplish the large-scale visual selection tasks. By another proposed novel mechanism for *temporary inhibition of return*, OADS can simulate the human saccading/attention behaviour to refixate/reattend interesting objects for further detailed inspection.

This thesis concludes that the proposed human-like visual selection solution – HOAF, which is inspired by psychophysical object-based attention theory and grouping-based competition, is particularly useful for machine vision. HOAF is a general and effective visual selection framework integrating object-based attention and attention-driven saccadic eye movements with biological plausibility and object-based hierarchical selectivity from coarse to fine in a space-time context.

Acknowledgements

Although my name alone is printed on the cover of this thesis, there are many people who, in one way or another, contribute in its realization. I would like to thank the following for their help in writing this thesis:

First of all, I would like to thank Professor Robert Fisher, my supervisor, for his truly outstanding guidance and continuous kind help during my PhD study.

I would also like to thank the School of Informatics, for funding my Scholarship, and British Overseas Research Scheme, for providing me with the ORS studentship.

I am grateful to Paul Crook and Rob Shipman for taking time reading and commenting on my thesis. Many thanks to Dr Herman Gomes for his generous help and suggestions on the log-polar foveated imaging technique and several chapters of this thesis. I thank all members of IPAB for helping me in many ways and my friends for various support and great fun over the years.

Finally, my most special thanks are to my family, for their love, encouragement, and supporting throughout these years.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(Yaoru Sun)

The work presented in this thesis is mainly based on the following papers:

Y. Sun and R. Fisher, "Hierarchical selectivity for object-based visual attention," Proc. 2nd Biologically Motivated Computer Vision Workshop (BMCV 2002) Tuebingen, Germany, November 2002, pp 427-438. Aka Springer LNCS 2525.

Y. Sun and R. Fisher, "Object-based attention for computer vision," *Artificial Intelligence*, Volume 146 (Issue 1), pp. 77-123, 2003.

Y. Sun, R. Fisher, F. Wang, and H. M. Gomes, "Object-based attention-driven saccadic eye movements," submitted to *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2003.

Table of Contents

1 Introduction **1**

1.1 Motivation 1

1.2 Proposed Solution and Original Achievements 5

1.3 Organisation of the Thesis 6

2 State of the Art of Visual Attention **10**

2.1 Introduction 10

2.2 Space-Based Attention 10

2.3 Object-Based Attention 15

2.4 Conventional Attention Models 18

2.4.1 Psychophysical Models of Attention 18

2.4.2 Machine Vision Models of Attention 21

2.5 Summary: Modelling Visual Attention for Machine Vision 23

3 Modelling Hierarchical Object-Based Attention **27**

3.1 Introduction 27

3.2 Theoretical Background from Psychophysics and Neuroscience 28

3.3 Hierarchical Object-based Attention Model (HOAM) 31

3.3.1 Overview of the Hierarchical Object-based Attention Framework (HOAF) 31

3.3.2 Fixation Image 34

3.3.3 Primary Feature Extraction 34

3.3.4 Grouping-Based Saliency Mapping 36

3.3.5 Competition Pool of Attention 47

3.3.6 Perceptual Grouping 52

4 Performance on Synthetic images and Natural Scenes **55**

4.1 Examination and Discussion 55

4.1.1 Performance in Synthetic Images 55
4.1.2 Performance in Natural Scenes 67
4.1.3 Improved Behaviour of Hierarchical Selectivity in Natural Scenes 82
4.2 Summary 86

5 State of the Art of Eye Movements 89
5.1 Introduction 89
5.2 Eye Movements and Visual Attention 90
5.2.1 Saccadic Eye Movements 90
5.2.2 Saccadic Eye Movements and Attention 91
5.2.3 Smooth Pursuit Movements 91
5.2.4 Pursuit Eye Movements and Attention 92
5.2.5 Vergence Movements 92
5.2.6 Vergence Movements and Attention 92
5.2.7 Conclusions of the Relationship Between Eye Movements and Attention 93
5.3 Conventional Machine Vision Models of Saccadic Eye Movements . . 93
5.4 Space Variant Sensing 95
5.4.1 Introduction 95
5.4.2 A Log-Polar Retina-Like Sensor 96
5.5 Summary 98

6 Modelling Object-Based Attention-Driven Saccadic Eye Movements 100
6.1 Introduction 100
6.2 Overview of OADS 102
6.3 Attention Window 108
6.4 Temporary Inhibition of Return (tIOR) 109
6.4.1 Spatio-Temporal Grouping Saliency Mapping 109
6.4.2 The Algorithm of tIOR 115
6.5 Attention-Driven Orienting (ADO) 117
6.6 Behaviour and Performance in Real-World Scenes 119
6.6.1 Implementation in Natural Scenes 119
6.6.2 Comparison with Other Work 124
6.7 Conclusion 135

7 Conclusions **136**

7.1 Contributions 137

7.2 Future Work 140

A Glossary **145**

B List of acronyms **147**

C List of Parameters **148**

Bibliography **149**

Chapter 1

Introduction

Traditional machine vision systems have used a space-based approach to model visual attention, resulting in malfunction in real-world visual environments. In this thesis, a highly effective and more biologically plausible approach is adopted that incorporates more recent psychophysical achievements on object-based visual attention to develop a novel Hierarchical Object-based Attention Framework (HOAF) for machine vision. The thesis describes the proposed approach and presents comprehensive experimental evidence on both psychological test images and real-world natural scenes to show that HOAF offers more effective and successful performance. In addition, HOAF shows more biologically-plausible visual behaviour on covert attention and overt saccading, especially on real-world natural scenes.

1.1 Motivation

It is well known that the primate visual system employs an attention mechanism to selectively process important information that is currently relevant to visual behaviours or visual tasks due to limited processing resources [106]. As a result, the balance between computing resource, time cost and fulfilling different visual tasks in normal, cluttered and dynamic environments can be efficiently dealt with. Visual attention selectivity can be either overt to drive and guide eye movements in order to pick up useful information over time [94, 43, 78] or covert, internally shifting the focus of attention from one locus to another without eye movements. Modelling visual attention is a challenging problem for machine vision [60] [105, p. 519-570]. Three closely-related fundamental questions are immediately identifiable:

1. How can the visual system know what information is important enough to cap-

ture attention?

Modern research on visual attention from psychophysical and neurophysiological experiments has found that there exists two ways by which information can be used to direct attention (see [148, 151] for reviews). One approach uses bottom-up information including basic features (e.g., colour, orientation, motion, depth, etc.), conjunctions of features (e.g., objects in 2D or 3D space), and even learned features. In this case, visually salient features [1] are mostly used to attract visual attention. A great number of attention models make use of "saliency" to direct attention [2, 4, 16, 75, 142]. However, saliency cannot always capture attention in a purely bottom-up fashion if attention is focused or directed elsewhere in advance [148, 151]. Thus it is necessary to recognize the importance of how attention is also controlled by top-down or goal-driven information relevant to current visual behaviours. The deployment of attention is determined by an interaction between bottom-up input and top-down attentional priming or setting [151].

2. How does the visual system know when and how to direct attention and choose important information rather than doing so at random times and by random selection?

   This is the paradox of intelligent selection of attention in visual systems. We would like to know whether selection happens earlier or later, to what extent visual processing is serial or parallel, and what interplay exists between these factors. A number of researchers have proposed two-stage models in which the preattentive stage performs independent detection or extraction of primary visual features automatically in parallel (without attention) and the second stage of attention processes the combination of primitive features by serially shifting the focus of attention to scan subsets of the incoming information available from the previous stage (see [148] for a review). This proposed strategy, however, conflicts with many modern psychophysical experiments that confirm that attention can arise from very early visual processing stages (e.g. feature detection) or arise from relatively late processing stages (e.g. object representation or recognition) in different circumstances in which parallel and serial processing reciprocally intertwine for effective performance of visual tasks [63, 84, 88]. Thus, this prob-

[1] A feature or stimulus differs from its immediate surround in some dimensions and the surround is reasonably homogeneous in those dimensions [29].

lem is far from well understood and requires further investigation.

3. Where is (are) the next potential target(s) of a visual attention shift? That is, how does attention know where to go and what to do next?

   There are two traditional assumptions in the literature attempting to account for this. The space-based attention theory holds that attention is allocated to a region of space, with processing of everything within this spatial window of attention like a spotlight, internal eye, or zoom-lens [38, 108, 137, 139]. Object-based attention theory argues that attention is actually directed to an object or a group of objects to process any properties of selected object(s) rather than regions of space [28, 70, 26, 122]. Some recent findings support a view that the two accounts are not mutually exclusive [32, 42, 63] and they may actually share common neural mechanisms in the parietal lobes [44]. Until now, few researchers have proposed attentional models that integrate space-based and object-based views (but see [87]). As suggested by S. E. Palmer in [105, p. 547-549], both hypotheses may be true to account for different processing levels respectively in the visual system and may be necessary to supply and interact at multiple processing levels for the coherent behaviours.

The above problems lead to a general question: How does visual attention work to perform effective selectivity? The dominant theory of visual attention is based on the hypothesis that attention works in spatial context like a "spotlight" or "zoom-lens", scanning the scene by shifting attention from one location to the next to limit processing to a variable size of space in the visual field. To date, there have been a number of attentional models for psychophysics or for machine vision that use this hypothesis. Most of them are derived from Treisman's Feature Integrated theory [137] which consists of separate low-level feature maps that are combined together by a spatial attention window operating on a master map or saliency map.

However, traditional attention models have only concentrated on mechanisms of visual attention based on selectivity by spatial locations. Thus they inherently lack mechanisms accounting for object-based visual selection (see [34, 35] and [105, p. 547-549] for reviews) and hence fail to work effectively in real-world natural scenes. Specifically, they have the following limitations:

1. Space-based attention models could not work in discontinuous or discrete space. A normal scene is usually cluttered: objects may overlap or share some common

properties. In this case attention may need to work in several discontinuous spatial regions at the same time so that an object can be selected at once.

2. Space-based attention could not account for object-based effects, such as "same-object advantages", "divided attention", "multiple object tracking", etc. [122]. Thus attentional models based on space-based attention do not work well in these cases.

3. The structure of an object may be very complex and hierarchical. In this case object-based hierarchical selectivity is required. Space-based attention can only select a spatial location or continuous region and does not work for this selection.

4. Human vision normally attends to meaningful information such as salient or interesting objects or regions when exploring a normal scene and rarely attends to nonsensical or empty locations. This is why object-based attention is much more effective than pure space-based attention in dealing with real-world visual environments.

We have known that attention and eye movements are often intertwined in human vision to carry out normal visual tasks. But they are different visual mechanisms. Attention is the primary mechanism of human visual selection, within which eye movements play the supporting roles and require attention to precede them to their goals [105, P. 570]. However, most of the traditional machine vision systems of overt selection (e.g., saccadic eye movements systems or foveal active vision systems) blur the difference between eye movements and visual (covert) attention by taking them as the same mechanism or completely ignoring the primary nature of attention in visual selectivity. Moveover, though some systems use an attention mechanism, they implement only space-based attention and none have exploited object-based selection. Therefore, they intrinsically lose general extendibility, robust search efficiency, biological-plausibility, hierarchical selectivity and the capability of selecting fewer empty or nonsense locations. Visual attention acts as the primary selection mechanism with eye movements playing an important supporting role but this has been rarely explored in machine vision systems. This, however, is one of the main novelties incorporated in the hierarchical object-based attention framework presented in this thesis.

In general, modelling visual attention for machine vision should achieve the following important objectives:

- Biological-plausibility;

- Object-based and space-based attention are integrated and work together at multiple selection levels;
- Visual attention and eye movements are modelled at different visual selection levels based on the central process of attention. Attention is the primary mechanism of visual selection with eye movements playing an important supporting role;
- Object-based hierarchical selectivity;
- Competition for visual attention in a space-time context;
- Satisfactory performance on both synthetic displays for the comparison with human attention behaviour and real-world natural scenes for applications.

1.2 Proposed Solution and Original Achievements

This thesis presents the proposed solution – a Hierarchical Object-based Attention Framework (HOAF) based on the grouping-based competition that attempts to address the limitations of conventional machine vision attention models and in particular, to achieve the above objectives for a machine vision selection system that can deal with complex visual selection tasks in real-world natural scenes.

HOAF is composed of object-based attention that is the primary mechanism of visual selection together with overt saccadic eye movements which is used to play a supporting role. These two kinds of mechanisms are modelled in separate levels of the integrated framework within which visual attention is used to select objects from coarse to fine scales and saccadic eye movements are guided by visual attention to help attention to achieve fine selection for the peripheral objects in the field of view. For this purpose, a Hierarchical Object-based Attention Model (HOAM) and an Object-based Attention-Driven Saccading model (OADS) are proposed to implement covert attentional selection and saccadic eye movements respectively, and to endow HOAF with object-based hierarchical selectivity.

This proposed solution – HOAF is comprehensively investigated through theoretical analysis, synthetic experiments and real-world natural scene applications. The resulting key innovations of HOAF are:

1. It achieves biologically-plausible object-based visual attention behaviour with visual hierarchical selectivity, built upon:

(a) grouping-based competition to integrate object-based visual attention with space-based visual attention;

(b) dynamic interaction between bottom-up visual grouping salience mapping and top-down attentional priming;

(c) object-based visual hierarchical selectivity in a space-time context;

(d) accurate and effective selection from far/coarse to near/fine under resolution-varying sensing;

2. It achieves human-like visual saccading behaviour, built upon:

(a) grouping-based competition for object-based visual attention/saccading;

(b) human retina-like foveal sensor for space variant representation;

(c) temporary inhibition of return for saccading/attention shifts in space-time;

(d) attention-driven saccadic eye movements for an integrated but distinguishable two-level covert and overt visual selection framework based on the mutual interaction and support between visual attention and saccading.

The HOAF presented in this thesis is believed to be the first implemented object-based selection solution with the capability of hierarchical covert and overt selectivity in the machine vision literature. Its biologically-plausible visual selection behaviour, successful performance, and object-based covert-overt selection are particularly useful when applied in real-world natural visual environments.

1.3 Organisation of the Thesis

This thesis is concerned with modelling object-based (covert) attentional selection and overt saccadic eye movements in an integrated hierarchical selectivity framework HOAF, to provide an effective human-like attention framework for machine vision. The first technical chapter (Chapter 3) provides an overview of HOAF and then focuses on visual attention model based on grouping-based competition to integrate object-based and space-based attention together. This is achieved by the proposed Hierarchical Object-based Attention Model (HOAM). Chapter 5 then shows how attention-driven saccadic eye movements working in a spatio-temporal context can be achieved in a biologically plausible way. The model of Object-based Attention-Driven Saccadic

movements (OADS) is proposed to reach this aim, which models visual (covert) attention and overt saccading as close-coupled two-level visual selections integrated within HOAF. Research background relevant to the above work is reviewed in chapter 2 and chapter 4 respectively. The last chapter sums up the overall work presented in this thesis.

The remainder of this chapter summarises the chapters listed above in more detail, as follows:

- **Chapter 2: State of the Art of Visual Attention**

  A comprehensive literature review for the research on visual covert attention in both psychophysics and machine vision areas is introduced in this chapter. This includes conventional space-based visual attention and recently developed object-based visual attention theories. It is then followed by a further analysis on the limitations of the space-based visual attention theory, the need for object-based attention and the advantages of integrating both theories together. This chapter also highlights a number of well-known space-based attention models developed in machine vision. The end of this chapter gives a summary and discusses in depth the proper approaches to model visual attention for machine vision.

- **Chapter 3: Modelling Hierarchical Object-Based Attention**

  The Hierarchical Object-based Attention Framework (HOAF) based on grouping competition is the proposed solution to model object-based visual attention for machine vision. HOAF is a hierarchical architecture within which object-based visual selection of covert attention and overt saccades are naturally combined together in a biologically-plausible way. The foundation of HOAF is the Hierarchical Object-based visual Attention Model (HOAM) which models visual (covert) attention and is presented in this chapter. The Object-based Attention-Driven Saccading model (OADS) built upon HOAM as an important supporting role of visual selection will be addressed in Chapter 5. These two models share some common mechanisms (e.g., low-level primary feature extraction, dynamic grouping-based salience mapping, etc.) and interact with one another in a two-level visual selection framework. Consequently, complex visual selection behaviour can be carried out effectively.

  Following the introduction and overview of the proposed attention framework HOAF, this chapter gives a formal description of the Hierarchical Object-based

Attention Model (HOAM) in detail, including primary features extraction (colours, intensity, and orientations), pyramidal feature maps construction, feature contrasts evaluation and grouping-based saliency mapping creation. The generation of top-down attentional priming and the mechanism of hierarchical selectivity are then proposed. In this work, dynamic grouping-based saliency mapping, grouping-based competition for visual attention and hierarchical selectivity are the keys to achieving the integration of object-based and space-based visual attention. Through the definition of "grouping", which is a hierarchically recursive structure formed by pixel(s), feature(s), object(s), region(s) or their grouping(s), object-based and location-based conceptions are merged. Because the structure and content of a grouping are formed dynamically depending upon space-variant resolution, the salience of a grouping varies correspondingly. This means that the structure of a grouping and its salience may vary with resolution and over time. Based on the dynamic grouping saliency mapping and interaction with top-down attentional priming, hierarchical selectivity drives visual attention to work from top-level groupings to sub-groupings and at the same time from coarse to fine and from far to near. As a result, the integrated visual selection of objects and space is implemented. To demonstrate the model's behaviour, a number of synthetic experiments designed based on the classical data from psychophysical research on visual attention, and real-world natural scenes are used to examine the model performance. The experimental results show that the model behaviour concurs with the main findings found in the psychological literature on object-based and space-based attention and achieves successful hierarchical selectivity of object and space.

- **Chapter 4: State of the Art of Eye Movements**

  Chapter 4 reviews the converging research achievements in psychophysics and machine vision on visual overt selection (i.e. eye movements). This review is mainly concerned with the relationship between visual covert attention and eye movements, saccadic eye movements, and space-variant sensing. In addition, a number of traditional saccading models are discussed. Most of them are not biologically-plausible as they completely neglect visual attention whereas others do not distinguish between visual attention and eye movements. With the support of abundant findings from the recent research on visual attention and eye movements, this chapter identifies the relationship between visual attention and

saccadic eye movements and gives a good starting point for the next chapter. Finally, the conventional approaches to build space variant sensors (e.g., log-polar retina-like sensor) are briefly reviewed.

- **Chapter 5: Modelling Object-Based Attention-Driven Saccadic Eye Movements**

  This chapter presents the proposed Object-based Attention-Driven Saccading model (OADS) built upon our previous work. In the model, object-based attention and saccadic eye movements are modelled at two levels in the integrated visual selection framework HOAF. Visual (covert) attention is the primary selection mechanism in HOAF. Saccadic eye movements are taken as a supporting role for visual (covert) attention to execute large-scale hierarchical selectivity by shifting the fovea from one fixated location to another. The novel innovations of this work arise from the proposed "Temporary Inhibition of Return" and "Attention-Driven Orienting" mechanisms. By using temporary inhibition of return, the model can simulate human attention/saccading behaviour to reattend to/fixate on interesting objects for further inspection. Through the attention-driven orienting mechanism, the foveal sensor fixates on the next destination guided by object-based attention. With this help, attention can select interesting objects located at the periphery of the whole field of view. Following the formal description of the proposed model, the experimental results obtained from real-world natural scenes are presented that examine the model performance and compare it with other traditional saccading models.

- **Chapter 6: Conclusions**

  In the final chapter, the conclusions of the work presented in the thesis are stated, including the highlighted results and original achievements. Unsolved problems and important future research raised by the work are then identified.

Chapter 2

State of the Art of Visual Attention

2.1 Introduction

Visual attention is a complex and extensive process or a set of processes that is difficult to define precisely. We can roughly define visual attention as the mechanism that allocates limited visual resources for processing selected aspects of the retinal image more fully than non-selected aspects [105, p. 532]. By using this intelligent visual selection, the visual system can flexibly explore the contents and layout of a complex visual field [89].

In the vast psychophysics literature concerning visual attention, there are two groups of theories regarding the underlying units of attentional selection: the traditional space-based theory and the recently developed object-based theory. The following two Sections 2.2 and 2.3 provide a brief review for these two areas. Section 2.4 discusses some well-known attention models developed in psychophysics and machine vision. Finally, Section 2.5 summarises some important issues related to computational approaches to modelling visual attention for machine vision.

2.2 Space-Based Attention

In the view of traditional space-based attention theories, visual attention selects continuous locations across space. For example, the "filter theory" [8] supposed that selection was to filter out non-selected information due to the limited perceptual resources; the "spotlight" theory [108] assumed that attention is like a spotlight to illuminate the focused location by moving along a path from one location to the next one through the operations of disengage-move-engage; "zoom-lens" theory [38] proposed that at-

tention is covertly directed to a region of space with the varying scope of the focus; the "Feature Integration Theory" (FIT) [137] suggested that attention serves to bind various properties of an object properly. All of these theories share a common, important characteristic: it is assumed that visual attention selectively focuses its available processing resources on whatever falls within a spatial region or even nothing at all.

The study of space-based visual attention is a long story and has many important findings. These findings are summarised below:

1. Properties of Attention [105, p. 532]:

1.1 *Capacity*: the amount of perceptual resources that are available for a given task or process, which can vary with alertness, motivation or time of day factors;

1.2 *Selectivity*: when the total capacity is fixed, the amount of attention can be allocated flexibly to some degree.

This suggests attention is dynamic, varying according to not only space distribution but also temporal load.

2. Functions of Attention [142]:

2.1 Selection of a region of interest in the visual field;

2.2 Selection of feature dimensions and values of interest;

2.3 Control of information flow through the network of neurons that constitute the visual system;

2.4 Shifting from one selected region to the next in time.

3. Goals or Benefits of Attention [82, p. 9-15]:

3.1 *Accurate Perceptual Judgements and Actions*: Attention can increase the accuracy of perceptual judgements by selecting information flow on the input side of cognitive processing, and can also increase the accuracy of actions on the output side of cognitive processing by selecting information flow in the organising and planning of both internal and external actions;

3.2 *Speeded Perceptual Judgements and Actions*: Attention increases the speed with which perceptual judgements and the planning/performance of actions take place;

3.3 *Maintenance of Mental Processing*: With attention, a perception or an action can be sustained for extended periods of time even when attention is not being driven by an expectation of a change in stimulus or a change in action;

3.4 *Controlling Order of Readout*: By attention, the different parts or details of the targets can be selectively perceived according to different visual tasks.

4. Properties of the Attended Area [82, p. 27-39]:

4.1 *The Boundaries of the Attended Area*: Evidence reveals that there exists a relatively sharp boundary between an attended object or area and its surroundings. According to the "spotlight" attention theory, it is plausible to assume that the boundary of the attended area falls between objects and seldom cuts across an object. However, the "zoom-lens" metaphor of attention suggested that the attended area has a high clarity at the center and a gradual decrease in clarity with distance from the center within a boundary of varied size and shape;

4.2 *The Variable Size of the Attended Area*: In the two-stage preattentive and attentive process models, attention can be set either widely or narrowly in a uniform distribution manner. The wide setting is appropriate for parallel processing of objects that "pop out" preattentively whereas the narrow setting is preferable for serial processing of objects constructed by conjunctions of attributes. The overall response time when spreading widely is greater than that when spreading narrowly. In the "zoom-lens" model, the distribution of attention may be adjusted continuously and attention operates to clarify details within the attended area;

4.3 *The Variable Intensity of the Attended Area*: The "spotlight" attention account supposed a constant intensity within a moveable attended area. The "zoom-lens" model assumed that the resolution of an attended area is varied but the overall intensity is constant. Some studies suggested an attended area of a particular size with varying amounts of resources and intensities of attention.

4.4. *Attending to One Connected Area at a Time*: "One-at-a-time" has been regarded in information-processing terms as a "bottleneck", or an attentional limitation in a dual-task performance. Recent research indicates that it does not always occur (although often) [106] and sufficient training and learning can improve the ability of simultaneous performance [130];

4.5 *The Movement or Shift of the Attended Area*: Space-based attention theories regarded attentional orienting as a covert analog of overt eye movements that move continuously across space. This metaphor is inspired by the studies that suggested some brain areas involved in eye movements (e.g., the superior colliculus and posterior parietal cortex) are also implicated in the processes of covert orienting.

4.6 *The Duration of Attention on the Attended Area*. There have been experiments to support a view that attention is a fleeting activity and can be sustained for a brief duration from a matter of milliseconds to seconds.

Most of metaphors of visual attention, such as "spotlight" or "zoom-lens", take into account some of the six properties of the attended area listed above, which imply attention enhances information flow in the target area, or inhibits the information flow in the surround, or both. The studies of how attention works still have a long way to go.

5. Control of Attention and Where to Go:

The control of attention concerns how visual attention is deployed or driven by the properties of objects and the goals of subjects. Since William James first introduced the concept in the well-known book "The Principle of Psychology" in 1890, there exists two major distinction about control attention, i.e., whether it is goal-driven, controlled in a top-down fashion in which the deployment of attention is the result of deliberation or intentions of attentional readiness ("active" in James' words); or stimuli-driven, controlled in a bottom-up fashion in which attention is captured by some salient attributes of objects that are not necessarily relevant to perceptual goals ("passive" in James' words) [151]. (Saliency requires two conditions: a stimulus that differs from its immediate surround in some dimension and a surround that is reasonably homogeneous in that dimension [29].) Later, Yantis concluded attention can be directed to locations in space "by a conscious and voluntary effort" and it can also be captured by abrupt onset and other stimulus events (the latter is faster and more potent than the former). But in both cases, top-down control plays a role [46, 147]. These results suggest that bottom-up and top-down mechanisms complement one another and more importantly, the deployment of attention in an image is determined by an interaction between the properties of the image and the observer's set of attentional goals [34]. There are two classical hypotheses in the literature about deployment of visual attention: space-based attention and object-based attention which will be discussed in the Section 2.3.

6. Directing Attention:

Attention can be covertly directed to a particular stimulus or to a location in the visual field. Posner [110, 111] proposed three fundamental components to covert orienting, which have been associated with different brain regions: disengagement of attention (parietal cortex), shifting of attention (superior colliculus) and engagement of the new location (lateral pulvinar of the thalamus). Eriksen [39] reported that an enhancement of processing at the cued location (a reduction in response time or an increase in accuracy) begins within 50 msec of a cue and continues to grow until it reaches asymptote about 200 msec after the cue, i.e. there doesn't appear to be abrupt, all-or-none switching, but instead a gradual buildup of attention at the cued location. This result, however, does not fit more recent results. Two different mechanisms can direct attention: one is stimulus-driven and another is goal-directed. Nakayama and Mackeben [96, 91] have argued there are at least two different dynamical forms of attention: one is transient, fast and involuntarily, i.e., the peripheral cue produces a quickly rising response and then falls to a lower asymptotic level (inhibition of return); another is slower, sustained under voluntary control, i.e., the central cue elicits a deliberate shift of attention by a monotonical drive towards an asymptote. They have also found that shifts in the transient form of attention can be unusually rapid when a temporal gap is placed between the disappearance of the fixation mask and the appearance of the target to be attended.

7. The Movements of (Covert) Attention:

It is now widely accepted [106, p. 80-88] that attention can shift from one location to another in the visual field without any concomitant eye movements. But whether attention moves in an analogous, continuous fashion, or whether shifts are abrupt without any actual movement still requires further exploration. Converging evidence from [81, 131] suggested that attention is "quantal rather than analog" and can be relocated independent of distance, and it can skip over an intervening obstacle without any time cost. This means that shifts of attention in space do not take time proportional to the spatial distance. Thus, the scheme serving to find the potential locations (or targets) may be parallel.

8. Spatio-Temporal Correlates of Attention:

The spatio-temporal correlates of attention, on one hand, can be observed in the activities of the overt attention associated with eye movements and, on the other hand,

can be measured by temporal characteristics of attentional deployment (e.g., early versus late selection or serial versus parallel processing, dwell time, switching attention, or movements of attention).

The study on the distinction between serial, focal attentional and parallel, preattentive processing has a long history [67, 98, 137, 139, 9]. The preattentive search is characterised by ("pop-out") search where the target is distinguished from distractors by at least one single basic feature (e.g., colour, size, and motion) with search times that are independent of the number of visual objects, while search for target(s) defined by conjunctions of features requires serial movements of focal attention. This division between serial and parallel processing is also linked to whether visual attention operates at an early or late stage (i.e. a lower-level processing stage such as feature detection or a higher-level processing stage such as object recognition). However, more recent research [140, 29, 147, 52] indicates that parallel and serial search properties can be obtained from the same set of target and nontarget stimuli. Parallel search may become serial if the target and distractors are made equally salient, and vice versa if the target is modified by attributes relevant to the task to increase its saliency. Thus the difference between serial and parallel mechanisms reflects the different control of attention in tasks. Parallel and serial search are accompanied by dynamic attention shifts [101, 103, 148]. Moreover, attentional selection may occur at late processing stages or operate at an early stage of processing under some conditions, mainly depending on whether the perceptual load is high or low [84, 88].

2.3 Object-Based Attention

In contrast to the space-based attention theories, the recent development of object-based attention argues that visual attention actually selects a perceptual object or group of objects rather than always selecting a continuous region of space [28]. In the view of object-based attention, the spatial location (of an object) is treated as one of the various properties (e.g., colour, shape, motion) of an object though in some cases location may have a higher ranking in processing (e.g., feature detection) than other properties. Object-attention concerns objecthood and object-based selection in a spatio-temporal context. Unlike space-based theories, spatial locations that do not contain any object are not considered in attentional selection. Research on object-based attention is still in development but has obtained a number of useful findings from psychophysics (e.g., [28], [35], [136]) and neuroscience (e.g., [23], [115]). The converging evidence for

object-based attention has been reviewed in [122]. The primary difference between object-based and space-based theories is the nature of the underlying unit of attentional selection.

There has been a rapidly increasing interest in object-based attention but research into useful systematic theories is still a very open area, especially practical models of object-based attention for real world applications. Several important issues should be addressed clearly in the further development of object-based attention:

- Early identification and segmentation of perceptual objects: this involves object representation, object-based selection between objects and within an object, and the relationship between object-based grouping and object-based attention;

- Neural substrate and related functionary mechanisms of object-based attention including both bottom-up and top-down factors that affect object-based biasing;

- The relationship between object-based and space-based attention: if it is true that they are actually not exclusive but operate at multiple selection levels in the visual system depending on visual tasks, how can they achieve the coherent selection by objects, features, locations, and their groupings? A recent study [44] shows that object-based and space-based attention share common neural mechanisms in the parietal lobes, temporal, and occipital cortices.

- Grouping/segmentation and object-based attention: this issue has been reviewed at length by Driver [27] who suggested segmentation and attention are mutually constrained and influenced. Without segmentation and grouping, object-based attention may lose its selection units. Note that the term "object" used to describe object-based attention is best to be thought as the term "proto-object" which is the result of segmentation processing and may also be hierarchically structured image regions rather than a normally experienced object (e.g., a solid-like apple) [27, 122].

  Perceptual grouping/organization is deeply intertwined with object-based attention, involving selective units, object-based advantages (e.g., same-object and same-grouping advantages, multiple object tracking, attending to parts, and inhibition of return, etc. [122]) and hierarchical selectivity (or multiple selective levels) by features, objects, or their hierarchically structured groupings. As suggested in [6], "object-based facilitation is a flexible and dynamic process operating at multiple spatial scales and over familiar and unfamiliar objects".

In Duncan's words, "The study of visual attention and perceptual organization must proceed together" [28]. However, one of the remaining questions is, when, where, and how the properties of an object, or elements of a grouping become a perceptive object or a grouping? Another question is, how the mutual impact between perceptual grouping and attention is evaluated or measured?

- Visual saliency and visual attention: we know that visual saliency can attract visual attention if the current top-down attentional setting is not fully loaded (or in other words, the current attention can be gained without top-down control) [151]. In this regard, we would like to know, what is the visual salience of a feature, an object, or a grouping? And what is the neural substrate to execute the saliency computation and judgement? How does visual saliency drive visual attention? The most important requirement to model visual attention in practice is how visual salience of a perceptual unit (whether a perceptual object or grouping) can be quantitatively measured, so that the saliency mapping of a visual field truly reflects its competitive situation for visual attention.

Research on the above issues is scarce in the literature of visual attention. In the following, a conceptual and asystematic theory concerned with object-based attention is introduced.

In the "Biased Competition" or "Integrated Competition" hypothesis proposed by Duncan and Desimone [23, 33], visual attention is taken as an emergent effect of competition in the visual information processing in multiple systems. In such a system activations working on different properties and actions of the same selected object become dominant through the competitive interactions biased by bottom-up and top-down influence relevant to the current behaviour. This account suggested three important principles for object-based visual attention: competition, behavioural relevance, and global integration of processing which favours the same selected object. This integrated competition hypothesis conflicts with the traditional views of visual attention (e.g., the binding role of visual attention [141]). Visual attention here is not thought of as a spotlight spatial process, or an external selective mechanism to create objects, but a widely distributed state which converges through the combined activation of multiple brain areas which collectively focus on the same selected object.

Duncan's hypothesis, especially the views of competition, integration, and behaviour dependance, is a useful guide in trying to develop a computational model of object-based attention, though it is far from complete and still requires a more system-

atic and theoretical description in order to build a practical model.

So far the above review is mainly focused on the research on visual attention which is the topic in this chapter. The research on eye movements that have close relationship with visual attention and are easily misunderstood by some people to be as the same as visual attention or to be able to directly generate visual selection without visual attention will be concerned in Chapter 4 and chapter 5.

2.4 Conventional Attention Models

There have been numerous attention models developed in the literature of visual attention. The review given below is just a subset of those that have been developed and focuses on well-known models which cover the majorities of approaches and common ideas found in psychophysical and machine vision modelling of visual attention.

2.4.1 Psychophysical Models of Attention

Feature Integrated Theory (FIT)

Both anatomical and physiological evidence support the hypothesis that the visual system divides input visual information into distinct subsystems that analyse and code different properties in various specialized areas. This raises a critical problem of how these dispersed representations are combined together into an unified perception, i.e., the binding problem.

Treisman's FIT model has been proposed to deal with this problem [137, 139, 140, 141]. FIT consists of a master map which codes locations of feature discontinuities in luminance, color, depth or motion, and a separate set of feature maps for processing information about the current spatial layout of the features. An attention window moves within a location map which selects the features attended to and temporarily excludes others from the feature maps, thus putting the "what" and "where" pathways together.

There are three spatially selective mechanisms used in FIT to solve the binding problem: selection by a spatial attention window, inhibition of location feature maps containing unwanted features, and top-down activation of the location containing the currently attended object. Figure 2.1 shows the model of FIT.

Treisman's model has made a number of experimental predictions that have been tested and confirmed in many experiments on space-based attention, such as conjunction search, texture segregation, illusory conjunctions and so forth [137, 138, 139, 140,

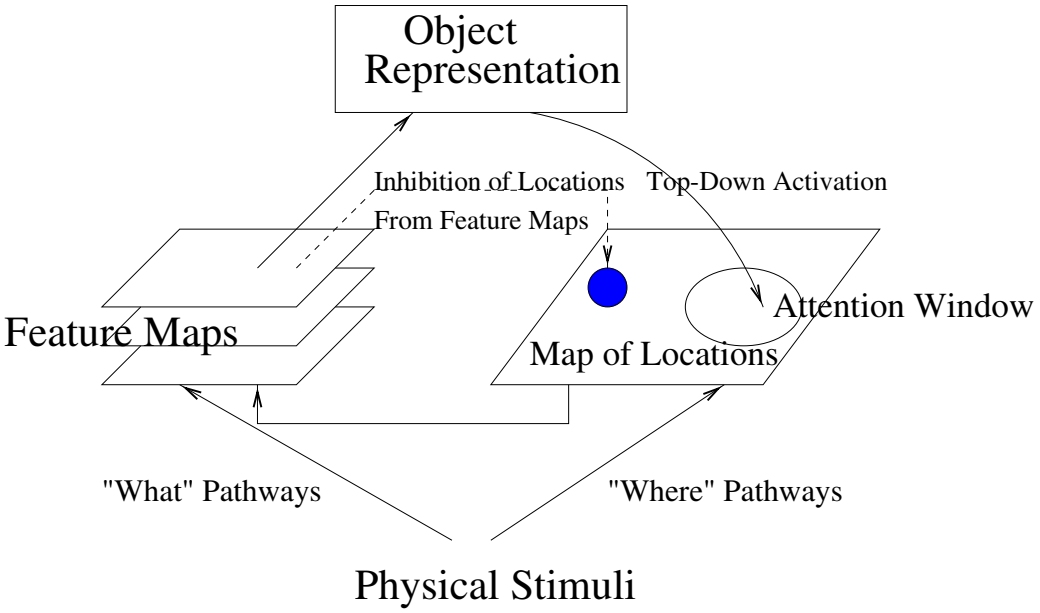

Figure 2.1: Treisman's FIT Model

141] (see [105, p. 556-560] for a review). Although her theory has been revised several times to fit new findings (e.g., the addition of inhibition mechanisms and the assumption that the attentional spotlight can act on feature dimensions to take account of the finding that many kinds of conjunctions (such as stereoscopic disparity and motion) can be detected in parallel), it does provide a general framework for understanding visual attention. Following her theory, a number of computational models of space-based attention have been developed. The main differences between these models are that they used different methods to construct and combine the low-level feature maps and to model the control mechanisms of attentional movements.

Dynamic Routing Circuits

This model was initially proposed by Anderson and Von Essen [1] and then further developed by Olshausen et al. [104]. The key objective of "dynamic routing circuits" is to make use of visual attention to route retinal information for translation-invariant pattern recognition. In the model, spatial attention is taken as a "window" shifting to different spatial locations in the retina. Activation from a retinal field passes through the attentional window before it reaches to the recognition systems. Local spatial relationships within the attentional window are preserved so that an attended object in the retina is represented within an object-centered reference frame.

Guided Search (GS)

The Guided Search model proposed by Wolfe [147] uses the well-known "Saliency Map" to account for visual search and focuses on simulating visual behaviour data. The model consists of two stages: a preattentive process which is built on a combination of bottom-up retinotopic-like feature maps and top-down commands for the computation of visual features; and an attention activation process which drives attention to the salient locations in a serial fashion. GS can account for the major findings of effective single feature and conjunction visual search. But the model did not consider the effects of grouping on visual search and most importantly, did not take object-based visual attention into account.

Search via Recursive Rejection (SERR)

The model SERR proposed by Humphreys and Müller [62] is built upon a hierarchical organization of networks and uses a Boltzmann-like machine activation function to simulate grouping effects on visual search with letter-like stimuli. It shows that visual search can be speeded up by rapidly rejecting distractor groupings and becomes increasingly ineffective as target-distractor grouping becomes stronger.

Selective Attention for Identification Model (SAIM)

The model SAIM (Heinke and Humphreys [56]) integrates "dynamic routing circuits" to achieve bottom-up invariant translation and a top-down knowledge network for object recognition via visual attention. It consists of three parallel processing networks: a contents network to map the contents of the visual field into the focus of attention; a selection network to determine the location of the mapped elements in the visual field; and a knowledge network to store object templates which directly influence selection. This model can perform location-based selection and object-based recognition using a competitive approach to attention.

Adaptive Resonance Theory (ART) Matching Rule

Grossberg [52, 53] proposed a new ART version to solve the attention-preattention (attention-perceptual grouping) and stability-plasticity dilemma problems. The latter concerns the fact that the brain is plastic and can rapidly learn new experiences without losing the stability that prevents catastrophic forgetting. The former concerns how laminar cortical circuits enable preattentive grouping processes to use some of the same

circuits that attentive mechanisms use, even before attentive mechanisms may come into play, in order to stabilize cortical development and learning. Grossberg suggested that both the bottom-up and top-down pathways contain adaptive weights, or long term memory traces, which may be modified by experience. The learned top-down expectations focus attention upon information that matches them. Then they select, synchronise, and amplify the activities of cells within the attentional focus, while suppressing the activities of irrelevant cells which could otherwise be incorporated into previously learned memories and thereby destabilise them. Such feedback resonance between bottom-up and top-down signals binding distributed information at multiple levels of brain processing into cortex-sensitive representations of objects and events. The bottom-up automatic activation, top-down priming, match and mismatch rules are together called the ART Matching Rule. In fact, the ART Matching Rule seems to be a later selection of attention, and is partly similar to Duncan's integrated competition hypothesis [33].

2.4.2 Machine Vision Models of Attention

The computational models of attention in machine vision are mainly inspired by psychophysical attention theories. In the early literature, many machine vision models aimed to develop foveal sensing mechanisms with overt eye movements and assumed these be identical to visual (covert) attention. However, visual attention and eye movements are in fact distinct visual selective mechanisms though they are usually intertwined to work together. This issue and related machine vision models will be addressed in detail in chapters 4 and 5. The review presented here focuses on the computational models that have modelled visual attention or applied visual attention mechanisms.

The attention model based on the "saliency map" was originally proposed by Koch and Ullman [75] and later implemented by Itti et al. [64]. In the model, the purpose of the saliency map is to combine the "salient" or "conspicuous" location information from each of the lower feature maps into a global measure to weight how different a given location is from its surround. This is used to guide selective attention. A winner-take-all (WTA) network implements the selection process by selecting the most conspicuous location in the saliency map. Attention is then directed to that location via a gating mechanism.

In general, this kind of saliency-based model contains three basic strategies: sev-

eral separate parallel feature maps to represent and code conspicuity within the visual field along particular dimensions (e.g., color, orientation, etc.); one (or more) saliency map(s) that combines different bottom-up inputs from feature maps; and a WTA (Winner-Take-All) mechanism that selects the most salient location in the saliency map for directing attention. This computational strategy has plausibly demonstrated some aspects of saliency-based attention and has also received some support from recent electrophysiological results [49, 119]. Due to its successful performance on many real images, this model has greatly influenced computer vision research. However, it is completely derived from the space-based attention theories and has inevitably been questioned by researchers working on object-based visual attention (e.g., [23, 33, 13]).

Tsotsos et al. [142] presented a selective tuning model for visual attention that used inhibition of irrelevant connections in a visual pyramid to realize spatial selection and a top-down WTA operation to perform attentional selection. In this model, visual spatial attention acts to optimise the search procedure and information processing.

In the model proposed by Clark et al. [16, 17], each task-specific feature detector is associated with a weight to signify the relative importance of that particular feature to the task and a WTA mechanism operates on the saliency map to drive spatial attention (as well as the triggering of saccades). In [51, 116], colour and stereo are used to filter images for attention focus candidates and to perform figure/ground separation.

Some researchers have used neural network approaches to model selective attention. In [4, 5], the saliency maps, which are derived from the residual error between the actual input and the expected input, are used to create the task-specific expectations for guiding the focus of attention. Kazanovich and Borisyuk proposed a neural network of phase oscillators along with a central oscillator (CO) as a global source of synchronization, and a group of peripheral oscillators (PO) for modelling visual attention [73]. Similar ideas have also been found in other work [20, 21, 80, 99, 100] and are supported by many biological investigations [80, 128, 143]. There are also some models of selective attention based on the mechanisms of gating or dynamic routing of information flow. This is generally achieved by dynamically modifying the connection strengths of neural networks [52, 62, 104, 112].

2.5 Summary: Modelling Visual Attention for Machine Vision

This chapter reviewed the major attention models in both psychophysics and machine vision. Until now, few machine vision and psychophysical models have incorporated object-based attention. The psychophysical models SEER [62] and SAIM [56] reviewed in Section 2.4.1 involved object-based selection but like the other psychological models, they are modelled to simulate behavioural data of human attention and can only be used to study attention in psychologically experimental paradigms such as stimulus-filtering and visual search tasks rather than in real-world visual environments. Although the machine vision models reviewed in the last section made good contributions to the implementation of location-based visual selection by modelling space-based attention, they all have the following serious problems:

1. They only applied the idea of space-based attention:
   As discussed in p. 3-4 of this thesis, the space-based computable models can only achieve location-based selection and can not perform the visual tasks that object-based attention accomplishes. This results in lots of odd, random, nonsense and bad selection in those machine vision models, especially when they were applied in real-world scenes in which object-based hierarchical selectivity is required (see Figure 6.12 in Section 5.6.2 for illustration).

2. Visual saliency is only evaluated in a uniform spatial context:
   Human vision uses spatial variant sensing and can freely re-explore interesting objects in a spatio-temporal context. In this way, visual saliency of an object varies with multiple resolutions and over time. In consequence it is required that the saliency mapping of the visual field and the inhibition of return mechanism for control of attention should be built in a spatio-temporal context for achieving human-like visual behaviour in machine vision.

3. The competition for attention is only location-based in a nontemporal context:
   Space-based attention makes use of location-based competition to perform location-based selection. However, visual objects compete for attention by their integration or ensemble effect rather than by their location features only [114]. Also, in order to simulate the human-like visual behaviour of revisiting the same object [94], the competition for attention should take object-based competition in a temporal context into account.

4. Spatial attention acts as same as saccadic eye movements to focus anywhere in the whole field of view:

   Human vision uses attention to acquire the detailed information of interesting objects in the visual field with high resolution surrounding the fovea without an eye movement. However, in order to investigate interesting objects located in the periphery of the whole field of view, attention needs overt saccadic eye movements over time to extend and improve the selection in a large-scale space [105, p. 519-571]. Therefore, the competition for attention exists anywhere but the exact attentional selection at a particular place requires the help of saccadic eye movements. The simulation of human-like visual selection needs overt orienting (eye movements) mechanisms to cooperate with the primary selection mechanism of visual attention. But it is important to make a clear distinction between visual attention and eye movements, which is often neglected in the previous attention models.

In order to solve the above problems in the existing work, a new approach is required to take the following aspects into account when building a biologically-plausible and highly effective attention system for machine vision:

1. Object-based Visual Attention:

   As discussed above, if an attention model is purely based on the space-based attention hypothesises, it is not complete nor biological plausible. Furthermore, it can not work successfully in normal real-world scenes in which object-based selection is required (see Introduction and Section 2.3 for related discussion in detail). A sound visual selection mechanism should not neglect object-based selection;

2. Distinction between (Covert) Attention Selection and Overt Foveal Eye Movements:

   As reviewed in this chapter and Chapter 4, visual attention covertly shifts in the visual field to select interesting objects when the fovea is fixated. Visual attention can perform visual selection without eye movements but eye movements require visual attention to function so as to assist attention to scrutinize the potential objects of visual selection in the periphery of the field of view [60]. Therefore, the shifts of attentional selection are clearly distinct from eye movements. Recently more and more active vision systems attempt to employ attentional mechanisms to help eye movements for their goal locating (e.g., [135, 3] but no research

of them has explored to integrate both of the shifts of (covert) attentional selection and eye movements in one system but distinguish them in two levels. A biologically-plausible vision system should consider foveal sensing together with visual attention but importantly make a clear distinction between them (this issue will be dealt with in detail in chapters 4 and 5).

3. Bottom-up and Top-down Interaction:
   Attention is controlled by the interaction of bottom-up and top-down influences. In other words, selective attention can not be automatically biased by pure bottom-up attractors if attention is deliberately directed elsewhere in advance. More importantly, this interaction, especially the top-down influence on attention, biases competition towards objects which are relevant to the current behaviour.

4. (Integrated) Competition:
   Attention is an emergent state or property arising from competitive interaction between units in the visual networks competing for prior selective processing due to limited visual resources. In this view, there is no single or specific "attention module" (in the sensorimotor network) which generates selectivity. Rather, attentional functions are distributed throughout multiple levels of information processing where competition takes place, as discussed in [33, 57].

5. Grouping-based Competition:
   Object-based attention holds that the underlying unit of attentional selection is an object or a grouping of objects as opposed to spatial locations. Therefore, grouping or segmentation is critical to object-based attention. Even to space-based attention, grouping or segmentation is also important since spatial attention does not always select a spatial location but a region in many cases, as suggested in the "zoom-lens" metaphor of space-based attention. In this thesis, grouping is not a simple equivalent to segmentation, but a key means to integrate both object-based and space-based attention together. This will be described in more detail in the next chapter.

6. Hierarchical Selectivity:
   This idea which is introduced and implemented in this thesis, derives from the following observations: 1) Attention can work at multiple processing levels to execute selectivity by features, objects, locations, or their groupings; 2) Competition for attentional selection permeates multiple levels of information process-

ing that results in hierarchical selectivity of attentional behaviour adapted to the current visual tasks; 3) Covert attention and overt eye movements, space-based attention and object-based attention work together at multiple levels to achieve structured selection of units (i.e., groupings in this thesis) relevant to the current visual behaviour.

7. Dynamics Over Space and Time:
   The dynamics of visual attention is embodied in the following elements: competition in space-time; the time course of deploying attention and of inhibition; the interaction between bottom-up and top-down attentional biasing, and saliency varying across space-time caused by eye movements.

In the remainder of this thesis a hierarchical object-based attention framework is proposed which aims to solve the problems of previous machine vision attention models by implementing the various aspects of an ideal attentional system listed above. These aspects have not been investigated in any conventional machine vision system so far implemented. This framework is the first implementation of an attention system integrating object-based and space-based attention together, which employs grouping-based competition to achieve object-based hierarchical selectivity of visual (covet) attention and attention-guided overt saccadic eye movements.

Chapter 3

Modelling Hierarchical Object-Based Attention

3.1 Introduction

Hierarchical Object-based Attention Framework (HOAF) (Figure 3.3) is a visual selection framework integrating object-based covert attention and overt saccadic eye movements. Two computational models (1) Hierarchical Object-based Attention Model (HOAM) and (2) Object-based Attention-Driven Saccading (OADS) are proposed to build this framework (as shown in Figure 3.1). The attention model (HOAM) is our first work version to model object-based attention in machine vision and is the primary component of HOAF. The saccading model OADS is our extended work for incorporating saccadic eye movements into the selection framework HOAF. These two models share common mechanisms (e.g., grouping-based competition) within HOAF and interact with each other to achieve object-based covert and overt hierarchical selectivity. The saccading model which is built upon the attention model and concerned with overt eye movements with visual attentional selection will be presented in Chapter 6. The current chapter concerns the overall architecture of the entire framework HOAF and its primary component's modelling for object-based attention. This presented work shows how a hierarchical object-based attention model can be developed through the approaches of grouping-based competition and hierarchical selectivity mechanisms. With the help of grouping-based competition, both object-based and space-based selection of attention are naturally linked together.

The remainder of the chapter is structured as follows. The next section addresses the theoretical inducements obtained from recent psychological achievements on vi-

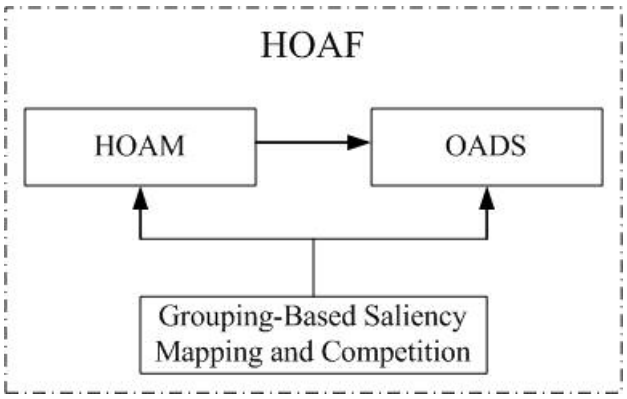

Figure 3.1: A diagram for HOAF within which the saccading model OADS is built upon the attention model HOAM.

sual attention for building the Hierarchical Object-based Attention Model (HOAM). The final section gives an overview of the proposed selection framework HOAF and then focuses on the technical description of its primary component – the object-based attention model HOAM.

3.2 Theoretical Background from Psychophysics and Neuroscience

Our work presented here for the Hierarchical Object-based Attention Model (HOAM) brings together many recent achievements and ideas from modern research on visual attention:

1. Object-Based Visual Attention such as the "Integrated Competition" theory [31, 32, 33, 23, 24] (reviewed in chapter 2):

   Our work has extended Duncan's Integrated Competition hypothesis so as to be capable of working in both object-based and space-based fields by replacing object-centered with grouping-centered (see one of the few psychophysical attentional models [10] and [53] for integrating object-based with space-based evidence).

   The concept of "grouping" here is defined as the underlying unit of attention selection and is used to link object-based and space-based attention together so as to obtain hierarchical selectivity within visual attention. A grouping is defined as a hierarchical structured unit segmented in the visual field by any acceptable

segmentation/grouping strategies, involving object(s), related features and locations, and their groupings (see [27, 63, 122] for further discussion on the issues related to the psychophysical concept "grouping"). In this way, a grouping here is similar to the traditional view of "grouping" as a visual segmentation or group but different and extended as accounting for a visual unit of attentional selection. Thus a grouping can be a point (e.g., a pixel in a scene), a feature (including a location), an object, a group of objects or features, or a region. Recent studies have shown that objects of attention can be defined on the basis of Gestalt grouping principles and familiarity [153].

Attention is an emergent behaviour of competitive interaction for visual selection in visual processing networks. At any given moment, enhanced responses to one grouping will decrease responses to other competitors. Once a grouping gains the dominance of selective attention, all other relevant processing in the visual system to this grouping and all sub-groupings belonging to this grouping share the same dominance. This is termed "integrated competition".

2. Bottom-up and Top-down Interaction of Visual Attention [148, 151]:

   The nature of attentional competition comes from the dynamic interaction between bottom-up visual salience and top-down attentional biasing or setting. That is, purely bottom-up or top-down driven information for attention can only bias the competition for selection process partly. In this case, salient visual groupings can capture attention quickly and automatically only if the current attention is not deliberately directed to other groupings or properties in advance.

3. Visual Saliency Map such as the saliency-based visual attention model [75, 64];

   Visual saliency is used to measure how different a location is from its surround. The more salient a location is, the more advantaged it is to compete with other locations for visual attention. The selection visual attention can be biased in a bottom-up manner by the saliency map which represents the spatial distribution of visual saliency of the field of view.

4. Spatial representation for Objects [63];

   Humphreys suggested that there exist two forms of spatial representations: within-object representation which codes elements as parts of a single object and links to the "what" pathway; between-object representation which codes elements independent objects and links to the "where" pathway. In addition, there is no

representation of space devoid of objects. Furthermore, spatial effects on visual selection are moderated by object representation. Object descriptions can be derived based on grouping between visual elements and grouped elements can be selected together.

5. Other Psychophysical Investigations on Object-based Attention [42, 31, 122].

   These studies revealed that spatial and object-based selection may be coupled by interactions between object and spatial processing systems. Spatial attention may affect object selection by biasing selection towards objects in the attended locations. Whereas object properties may affect spatial selection to make spatial attention be locked onto objects.

One of the novel mechanisms proposed in our work to serve the purpose of integrating object-based and space-based attention together is the grouping-based competition for visual attention based on the interaction between bottom-up visual saliency mapping and top-down attentional priming. The early visual features of a scene (colours, intensity, and orientations) are extracted by multiresolution pyramids. The visual salience of points, objects, regions, or groupings at different resolutions is calculated respectively on the pyramids to build up a grouping-based salience pyramid – a basis of the purely bottom-up attention competition among various visual inputs. The competition for visual attention is modulated by the interaction between such bottom-up visual saliency and the top-down attentional setting which is decomposed into positive priming, negative priming, free, and occupied cases.

Another novel mechanism proposed in our work serving to guide visual attentional movements is hierarchical selectivity, which can be regarded as a kind of multiple selectivity [105, p. 547-554] integrating attentional selection by spatial locations, visual features and their complex conjunctions (e.g. objects or groupings). The competition for attention takes place first from the most coarse level on multiresolution pyramids, then gradually to the finer level, as well as from coarser groupings to finer groupings within and between groupings and resolutions. This kind of behaviour of visual hierarchical selectivity derived from our work is similar to the human's behaviour of visual attention (see [6, 122] and [105, p. 594] for the reviews of visual attention selectivity by grouping, spatial regions, objects, parts and properties of an object) and is thus biologically plausible. Hierarchical selectivity can be used to guide the attentional movements shifting from one locus of attention to another under a multiscale transformation.

The work presented here implies that visual attention can directly select a continuous area of space, discrete object(s), feature(s), point(s), or their grouping. The space-based and object-based attentional selectivity are either cooperative or independent of each other for effective selective acts according to the current visual situation and tasks. This strategy is especially useful for machine vision. For example, space-based selection can be applied to region segmentation whereas object-based selection can be used for object recognition or fine analysis.

Our work explores the first machine-vision implementation of an object-based visual attention system integrating space-based attention and hierarchical selectivity.

3.3 Hierarchical Object-based Attention Model (HOAM)

This section presents the Hierarchical Object-based Attention Model (HOAM) – the foundation and primary component of the Hierarchical Object-based Attention Framework (HOAF) in detail. In the following, we first introduce an overview of the entire architecture of the framework HOAF.

3.3.1 Overview of the Hierarchical Object-based Attention Framework (HOAF)

The architecture of the proposed object-based selective solution HOAF is schematically illustrated in Figure 3.3. HOAF consists of the following modules: retina-like sensor, attention window, feature primary feature extraction, feature maps (colour, intensity, and orientation pyramids), grouping saliency mapping, temporary inhibition of return (or short-term memory), and competition pool of attention. HOAF consists of two close-coupled models: (1) the Hierarchical Object-based Attention Model (HOAM) for modelling object-based visual attention and (2) the Object-based Attention-Driven Saccading model based on the extended work of HOAM for incorporating saccadic eye movements to assist visual attention to achieve more flexible visual selection. The object-based attention model presented in this chapter is the kernel of HOAF and responsible for triggering both attentional selection and saccadic eye movements, which is mainly composed of primary feature extraction, feature maps building, grouping saliency mapping, and the competition pool of attention.

As shown in Figure 3.2, the proposed model firstly extracts primary features (colours, intensity, and orientations) from one fixated image (i.e., the fovea is fixated at a lo-

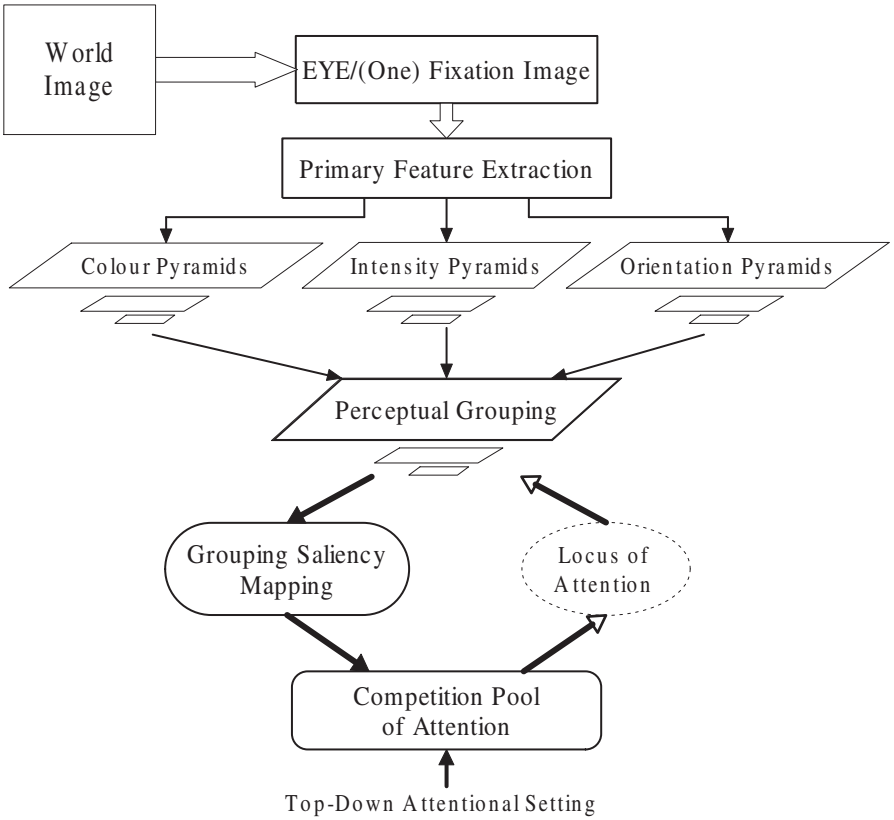

Figure 3.2: The schematic description of the Hierarchical Object-based Attention model HOAM.

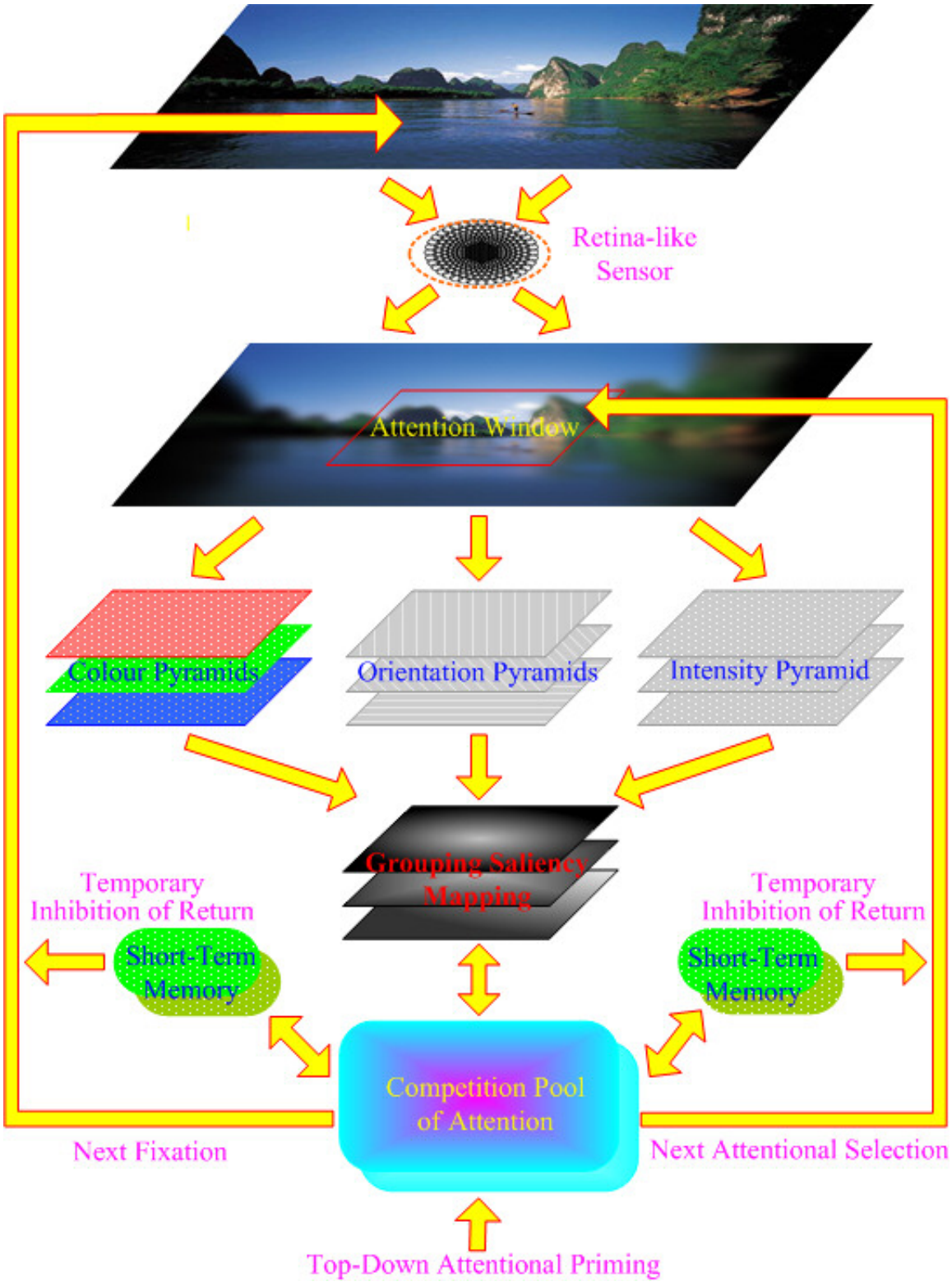

Figure 3.3: The architecture of HOAF

cation) sampled from a given scene, and then builds pyramid-like feature maps by overcomplete steerable filters. After perceptual grouping preprocessing, the bottom-up saliency mapping of various groupings is created via the grouping-based saliency computation. The obtained saliency mapping is varied with the dynamic competition among groupings dynamically created at different resolutions during attentional shifts. The results produced from this stage are fed to the attention competition pool where all groupings compete against each other to preferentially gain the selective attention. This kind of competition first occurs between the top level groupings at the coarsest resolution and then shifts to either the winner's sub-groupings or the unattended top level groupings from coarse resolution to finer resolution guided by the "hierarchical selectivity" mechanism (see Figure 3.8 for the illustration). Meanwhile, the grouping based competition is a dynamic interaction between bottom-up saliency and the top-down attentional setting. The rules of winner-take-all and inhibition of return are applied here to ensure the winner benefits and prevent attention from returning to the previously attended groupings. The detailed description of each module and related issues of this model is given in the following sections.

3.3.2 Fixation Image

At any moment, a fixation image or foveated image, which is a transformation of the world image into retinal image at each fixation point, is obtained by simulating the functional mapping of resolution decreasing from the fovea to the periphery of the retina. The following modules involved in the model operate on a given fixation image created from the original scene by a gaze sensor at that moment. In this chapter a fixation image is assumed to be produced by a resolution-uniform sensor sampling on the input scene and this assumption is well accepted by most of attention models. Chapter 5 will describe a space variant sensor (a human retina-like sensor) to create the resolution-varying foveated images for the model of saccadic eye movements and show how object-based attention is integrated with overt multiple saccading fixations in a spatio-temporal context.

3.3.3 Primary Feature Extraction

The colour input image (i.e., a fixation image) is decomposed into sets of multiscale feature maps via overcomplete steerable pyramid filters [50], to generate four colour, one intensity and four (or eight) orientation pyramids [64]. Suppose that F is the input

image, with r, g, b being the red, green, and blue colour components of F. An intensity image $I(p_{ij})$ is created by:

$$I(p_{ij}) = [r(p_{ij}) + g(p_{ij}) + b(p_{ij})]/3 \tag{3.1}$$

where p_{ij} is a point of F, $i \in [1 \dots n]$, $j \in [1 \dots m]$, $n \times m$ is the size of the image.

Then, four colour channels R (red), G (green), B (blue), and Y (yellow) are obtained as in [64] (negative values are set to zero):

$$\begin{aligned}
R(p_{ij}) &= r(p_{ij}) - [g(p_{ij}) + b(p_{ij})]/2 \\
G(p_{ij}) &= g(p_{ij}) - [r(p_{ij}) + b(p_{ij})]/2 \\
B(p_{ij}) &= b(p_{ij}) - [r(p_{ij}) + g(p_{ij})]/2 \\
Y(p_{ij}) &= [r(p_{ij}) + g(p_{ij})]/2 - \left|r(p_{ij}) - g(p_{ij})\right|/2 - b(p_{ij})
\end{aligned} \tag{3.2}$$

The idea behind the above computation is similar to many other attention models (e.g., [95, 64]). The rgb colour space is used to produce four colour channels for further building a double-opponent (RG and BY) colour space so that colour contrast computation can be not only easily achieved by a simple and linear transformation but is also biologically-plausible as we have known the primate visual cortex broadly uses double-opponent colour channels for colour tuning and contrast measurement [36].

Let W_{lpf}, $W_{bpf}(\lambda;\theta)$ be 2D Gaussian and orientated Gabor steerable filters respectively. With these filters acting on the five I, R, G, B, and Y channels (see [50, 64] for more details), we can construct intensity, colour (red, green, blue, and yellow) and orientation pyramids:

$$I_{\lambda+1} = W_{lpf}^T \cdot W_{lpf} \cdot I_{\lambda;} \;\; I_0 = I \tag{3.3}$$

$$\begin{aligned}
R_{\lambda+1} &= W_{lpf}^T \cdot W_{lpf} \cdot R_{\lambda;} \;\; R_0 = R \\
G_{\lambda+1} &= W_{lpf}^T \cdot W_{lpf} \cdot G_{\lambda;} \;\; G_0 = G \\
B_{\lambda+1} &= W_{lpf}^T \cdot W_{lpf} \cdot B_{\lambda;} \;\; B_0 = B \\
Y_{\lambda+1} &= W_{lpf}^T \cdot W_{lpf} \cdot Y_{\lambda;} \;\; Y_0 = Y
\end{aligned} \tag{3.4}$$

$$O_{\lambda}(\theta) = W_{bpf}(\lambda;\theta) \cdot I \tag{3.5}$$

where $\lambda \in [1 \dots l]$ is the pyramid's scale, $\theta \in [0^0, 45^0, 90^0, 135^0]$ or $[0^0, 22.5^0, 45^0, 67.5^0, 90^0, 112.5^0, 135^0, 157.5^0]$ (we used both orientation sets for different experiment environments but the first is the general one) is the preferred orientation, and "$\cdot$" is the convolution operator. The Anderson kernel used for W_{lpf} is $(1/16, 1/4, 3/8, 1/4, 1/16)$.

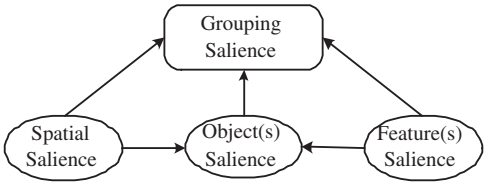

Figure 3.4: Diagram of grouping salience

The Gabor filter comes from modulating the related Lapacian pyramids with a set of oriented sine waves, then being followed by a low-pass filter operation, and finally taking the modulus (see [50] for these two filters in detail).

3.3.4 Grouping-Based Saliency Mapping

Competition for visual attention based on groupings is the bridge to achieve object-based attention and integrate space-based attention in HOAM. In this approach, a grouping is the primary perceptual unit upon which attentional selection operates. The term "grouping" is a common concept in the long research history of perceptual grouping by the Gestaltists (see [105, p. 257-266] for a review). Saliency mapping is evaluated based on groupings because "grouping" itself has already embedded "object" and "space". This usage constitutes a fundamental difference to most of the previous computable models of space-based attention. As already discussed in Section 3.2, the concept of grouping defined here can be regarded as an extension of the concept "perceptual grouping" in the traditional Gestalt story. A grouping is a hierarchical structure of objects and space. In this sense, a grouping may be a point, an object, a region, or a hierarchical structure of groupings. In this thesis, it is assumed that a given scene at each scale has already been segmented into groupings according to the Gestalt principles (or other grouping approaches). Clearly, the theory presented here for grouping-based saliency mapping is independent of whatever a technical approach used for practical segmentation which thus is not concerned here.

The salience of a grouping is a function of all saliency contributions coming from the components within the grouping working together to compete with their common competitors and competing with each other. This notion covers two issues. One issue is the relationship of spatial location, objects, and features to the grouping they belong to, as shown in Figure 3.4. The figure shows that grouping salience is computed from its components of spatial location, feature(s), and/or object(s). The other issue is the

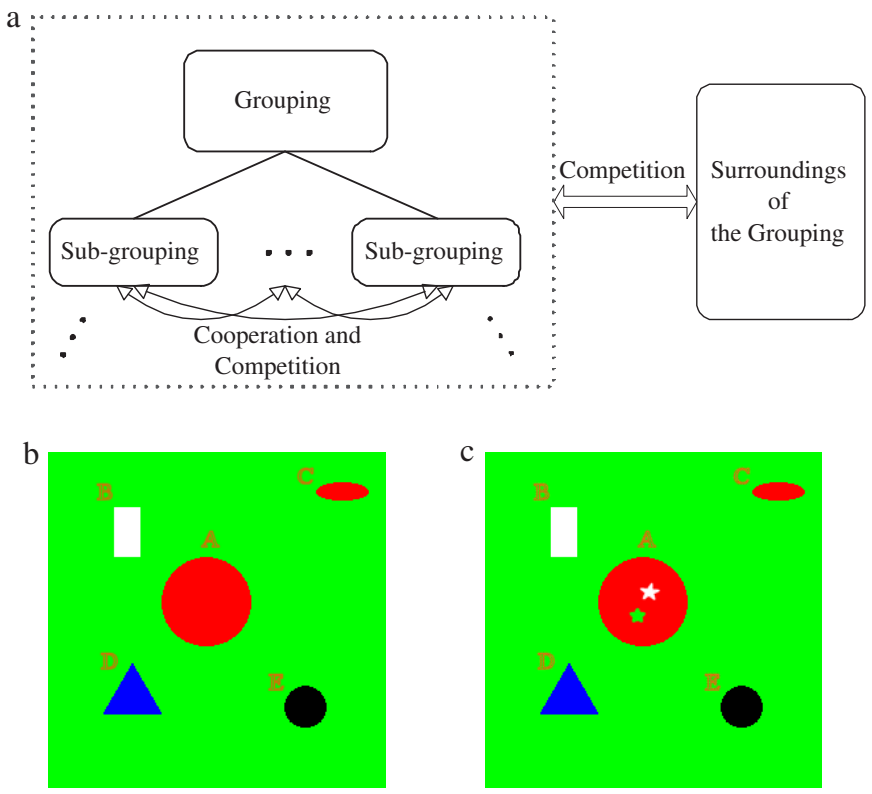

Figure 3.5: An example of grouping salience

competition between a grouping and its surroundings by cooperation and the competition between its components. The effect of a competition between two competitors may either enhance or suppress their salience according to their contrast properties (see Figure 3.5 (a)). Two simple examples are given in Figure 3.5 (b) and (c).

Suppose the red circle (grouping A) is the target. We want to calculate its salience. Its surroundings consist of four groupings (B, C, D, E) and other background points (green pixels). In Figure 3.5 (b), all red pixels within grouping A work together to enhance the grouping salience by feature contrast to compete with the surroundings. Along with this global competition, local competitions among pixels within grouping A also produce a negatively enhanced effect on the grouping salience due to the same features of these pixels. In Figure 3.5 (c), the green star sub-grouping in grouping A brings a suppressive effect on the total A grouping salience when it competes with green pixels in the background but a enhanced effect when it competes with the non-green groupings and pixels with A and elsewhere. The final salience of grouping A depends on the competitive effects brought by all of the components within A (including red pixels, white star and green star).

Based on the above considerations, the contrast between any two points is the primitive operation in the computation of grouping salience. However, we are not claiming that the salience computation theory introduced below is complete. Our research is concerned with salience deriving from the colour, intensity, orientation, and distance factors only. Many other factors affecting salience are not included here, such as motion, shape, size, depth and the like (see [148] for the related issues in visual search). One unconsidered factor about relative size difference between groupings will be discussed later (see Section 3.3.4.3).

The salience of a grouping is calculated by combining the colour, intensity, and orientation salience of the components of the grouping. Due to the close relationship between the chromatic opponent-colour channels and the achromatic (white-black) channel in the visual perception and contrast process [15, 144], we calculate colour and intensity salience together.

Suppose $\Re$ is any given grouping at the current resolution scale λ at time t, Θ is the surroundings of $\Re$. Let $\forall\Re_i \in \Re$ represent a sub-grouping of $\Re$, $\forall\Re_j \in (\Re \cup \Theta)$ represent a sub-grouping of $\Re$ or a grouping within Θ and $i \neq j$, we calculate the colour-intensity salience S_{CI} and orientation salience S_O of $\Re_i$ by:

$$S_{CI}(\Re_i;\lambda;t) = f_{CI}(\Re_i;\{\Re_j\};\lambda;t)$$

$$S_O(\Re_i;\lambda;t) = f_O(\Re_i;\{\Re_j\};\lambda;t) \tag{3.6}$$

where f_{CI}, f_O are the calculating functions of colour-intensity, orientation salience between $\Re_i$ and $\Re_j$ respectively. The salience S of grouping $\Re$ is given as:

$$S(\Re_i;\lambda;t) = \Gamma[S_{CI}(\Re_i;\lambda;t);S_O(\Re_i;\lambda;t)]$$

$$S(\Re;\lambda;t) = \Psi[S(\Re_i;\lambda;t)] \tag{3.7}$$

where Γ, Ψ are normalization and integration functions respectively. These functions are further defined in detail by Eq. 3.14, 3.20 and 3.21.

This saliency computation is built upon the principle of contrast between a centre and its local and global surround while the neighourhood distribution and the dynamics are also considered as covert attention occurs. As pointed out in [49], most (stable) objects in a normal environment are not intrinsically salient but can become salient if they are behaviourally significant. The normal scene has a hierarchical structure, thus features may not always have the same salience when viewed in extended regions or larger contexts. In other words, the salient difference among objects or features may

change over time, or as background or the context of the scene changes. The saliency computation is a complex and difficult problem. Until now few research studies in the field of attention in machine vision have dealt with it (however, see [64, 65, 126] for some discussion related to spatial saliency map). From our point of view, visual saliency arises from the competition between different groupings and between a grouping and its surroundings.

For simplicity in formulas, all computations below are defined for a given current time and resolution scale. The salience computation at other times and spatial scales is similar because the salience of a grouping is decided only by the current constitution of the grouping and its surroundings. Thus the changing of salience over time (salience dynamics) of a grouping depends upon the varying of the grouping's current constitution and surroundings over time. That is, the same computation rules are used for any time and scale when the segmentation of groupings at that time and scale is given. In this way, the full details of the computable approach are given below.

3.3.4.1 Colour and Intensity Salience

Assume x, y are two arbitrary pixels in a grouping $\Re$ on level λ of pyramids of colours, intensity, and orientations. Then, the properties of x and y can be denoted by a tensor composed of a 4-dimension colour vector, a 1-dimension achromatic intensity vector, and a 4-dimension orientation vector. For example, pixel x = $(\{R_{x,\lambda,\Re}, G_{x,\lambda,\Re}, B_{x,\lambda,\Re}, Y_{x,\lambda,\Re}\}, \{I_{x,\lambda,\Re}\}, \{O_{x,\lambda,\Re}(\theta_1), O_{x,\lambda,\Re}(\theta_2), O_{x,\lambda,\Re}(\theta_3), O_{x,\lambda,\Re}(\theta_4)\})$. We suppose here all calculations are within a given group on a given pyramid level because calculations on other pyramid levels are similar. The issue about whether and how to combine multiple scales will be discussed later in Section 3.3.4.4. Hence, the subscripts $\Re$ and λ will be generally omitted in the following description. We first compute the property contrast between pixels x and y. Let RG and BY be the two colour "double-opponent channels" of red-green/green-red and blue-yellow/yellow-blue [36, 68], so we have:

$$RG(x,y) = |(R_x - G_x) - (R_y - G_y)|/2$$

$$BY(x,y) = |(B_x - Y_x) - (B_y - Y_y)|/2 \tag{3.8}$$

The colour chromatic contrast ΔC between x and y is calculated as:

$$\Delta C(x,y) = \sqrt{\eta_{RG}^2 RG^2(x,y) + \eta_{BY}^2 BY^2(x,y)} \tag{3.9}$$

where η_{RG} and η_{BY} are the weighting parameters. In the experiments of this chapter, we set them as:

$$\eta_{RG} = \frac{R_x + R_y + G_x + G_y}{R_x + R_y + G_x + G_y + B_x + B_y + Y_x + Y_y}$$

$$\eta_{BY} = \frac{2\sqrt{B_x^2 + B_y^2 + Y_x^2 + Y_y^2}}{3 \times 255} \tag{3.10}$$

where the 255 parameter is used because the representations of colour and intensity have here the maximum value 255. The weights η_{RG} and η_{BY} can be optimized further according to more colour discrimination experiments or references in the colour research literature. The results produced by setting η_{RG} and η_{BY} as those in Eq. 3.10 are very close to L*u*v* (see [97, 113] for related issues). We obtain equal maximal contrasts between opponent colours such as red and green, blue and yellow, or white and black. The contrasts between other colours are also reasonable. For example, it is acceptable that the colour contrast between yellow and black is greater than yellow and white, etc. (see [71, 97, 150] for more discussion). All values of colour-intensity contrasts between x and y fall into the range $[0 \ldots 255]$.

The intensity contrast between the two pixels x and y is:

$$\Delta I(x,y) = |I(x) - I(y)| \tag{3.11}$$

So, the formula for calculating the salience $S_{CI}(x,y)$ of colour-intensity between x and y is:

$$S_{CI}(x,y) = \sqrt{\alpha \Delta C(x,y)^2 + \beta \Delta I(x,y)^2} \tag{3.12}$$

where α and β are weighting coefficients and we here set them to 1. The calculation here is simply inspired by evaluating the distance between two points in a colour-intensity contrast space and also making the combined colour-intensity salience increase with either colour or intensity contrast increases. By setting $\alpha = \beta = 1$, this simple calculation will make the combined salience have equally weighted colour and intensity contrast.

Suppose d_{gauss} is the Gaussian weighting function which evaluates the contrast effect along the distance between x and y. This Gaussian weighting is defined as:

$$d_{gauss}(x,y) = (1 - \frac{||x-y||}{\hat{n}-1}) e^{-\frac{1}{2\sigma^2}||x-y||^2} \tag{3.13}$$

with the scale σ and distance $||x-y||$. In the experiments of this chapter, the Gaussian scale σ is set to $\hat{n}/\rho$ where $\hat{n}$ is the maximum of the width and length of the

feature maps on the current pyramid level λ. ρ is a positive integer and generally $1/\rho$ may be set to a percentage of $\hat{n}$, such as 2%, 4%, 5%, or 20%, 25%, 50%, etc. The greater ρ is, the smaller the radius between the neighborhood and its surrounding center is. In this way, the Gaussian weighting guarantees competition throughout the attention window but the strength varies with distance. This function produces strong local competition between short-range neighbours and weak competition between long-range neighbours. Such similar effects of attention competition have been found in visual cortex [24]. Research on cortico-cortical connections shows that inhibition from the surround of the same stimulus properties as the center is strongest [127]. The distance $||x-y||$ can be the Euclidean distance but we prefer a chessboard distance: $||x-y|| = MAX(|i-h|,|j-k|)$, (i,j), (h,k) are the coordinates of x, y on the current pyramid level. *MAX* denotes the maximizing operator. The reason for selecting the chessboard distance is that the neighbours within the same 8-adjacency neighbourhood have equal distance effects on their common center so that the contrast between a location and its surround can be evaluated in the approximately circular manner.

Let $\mathcal{NH}_{CI}$ be the neighbourhood surrounding x, $y_i \subset \mathcal{NH}_{CI}$ $(i = 1 \ldots n \times m - 1)$ be a neighbour, $n \times m$ be the size of a pyramid level. We use the following formula to calculate the colour-intensity salience of x:

$$S_{CI}(x) = \frac{\sum_{i=1}^{n\times m-1} S_{CI}(x,y_i) \cdot d_{gauss}(x,y_i)}{\sum_{i=1}^{n\times m-1} d_{gauss}(x,y_i)} \tag{3.14}$$

3.3.4.2 Orientation Salience

Define $\overline{\theta}_{x,y}$ as the orientation difference between pixels x and y. Let $u_x(\theta)$ and $u_y(\phi)$ be the orientation vectors of x and y in the current orientation pyramid respectively. Note that u, θ, and ϕ themselves all consist of multiple components. For example, $u_x(\theta) = [u_x(0), u_x(\frac{\pi}{4}), u_x(\frac{\pi}{2}), u_x(\frac{3\pi}{4})]$, if we have four preferred orientations. We define the orientation salience $C_O(x,y)$ of x to y as:

$$C_O(x,y) = d_{gauss}(x,y)\sin(\overline{\theta}_{x,y}) \tag{3.15}$$

where d_{gauss} has already been defined in Eq. 3.13. A major reason to select a sinusoid function for orientation contrast is that this function is a nonlinear and monotonically increasing function from 0 to 1 over the range $[0, \frac{\pi}{2}]$ and symmetric in $[0,\pi]$. Nothdurft has suggested that the salience of pop-out targets has a nonlinear (enhanced) character

from threshold and saturation effects with increasing orientation contrast from 0 to $\frac{\pi}{2}$ [102]. If u_x and u_y have orientation strengths at all orientations, then the general calculation for $\overline{\theta}_{x,y}$ can be given by:

$$\overline{\theta}_{x,y} = \frac{\int_0^{\pi} \phi \left[\int_0^{\pi} u_x(\theta) u_y((\theta+\phi) \mod \pi) d\theta \right] d\phi}{\int\int_0^{\pi} u_x(\theta) u_y((\theta+\phi) \mod \pi) d\theta d\phi} \tag{3.16}$$

For practical computation in this chapter, we give the following discrete form for $\overline{\theta}_{x,y}$:

$$\overline{\theta}_{x,y} = \frac{\sum_{j=0}^{\zeta-1} j\varphi \sum_{i=0}^{\zeta-1} u_x(i\varphi) u_y((i\varphi + j\varphi) \mod \pi)}{\sum_{j=0}^{\zeta-1} \sum_{i=0}^{\zeta-1} u_x(i\varphi) u_y((i\varphi + j\varphi) \mod \pi)} \tag{3.17}$$

where **mod** is the standard modulus operator, ζ is the number of orientation pyramids or preferred orientations, $\varphi = \pi/\zeta$. When ζ is 4 or 8, φ is $\pi/4$ or $\pi/8$.

The salience computation for orientation is more complicated than for colour-intensity. It is most important to take into account the homogeneity/heterogeneity of the neighbourhood of each point which is currently taken as a center for center-surround calculation. Psychophysical findings show that "pop-out" is closely related to the distribution of orientations in the local neighbourhood [82, 106, 137, 140, 148]. Aiming at a practical computation of orientation salience, further considerations of "center-surround" operations are provided as follows.

Let y_i ($i = 1 \ldots n_k$, n_k is the number of neighbours in the k-th neighbourhood) be a neighbour in the distance k or k-th neighbourhood $\mathcal{NH}_O(k)$ surrounding x. It is clear that the distance 1 or first neighbourhood of x has 8 closest neighbours surrounding x, and that the distance k neighbourhood has $8k$ neighbours. A boundary check must be applied to ensure all data comes from within the current image layer. Then the average orientation contrast of x to its k-th neighbourhood is:

$$\overline{C}_O(x, \mathcal{NH}_O(k)) = \frac{1}{n_k} \sum_{y_{i \in \mathcal{NH}_O(k)}} C_O(x, y_i) \tag{3.18}$$

Suppose n_0 is the number of different directions within $\mathcal{NH}_O(k)$, we have $\omega_k = n_0 - 1$. This is used for checking and evaluating how heterogeneous the orientations are in the neighbourhood of x. n_0 can be obtained by a simple method: set $n_0 = 0$; then $n_0 = n_0 + 1$ if the orientations of any y_i and y_{i+1} are different from the orientation vector's members that have the maximum values respectively in their orientation (vector) maps.

The same set of histograms above is used to evaluate the orientation homogeneity of the whole surround of x. Let w_{ijk} be y_i's value on the orientation (θ_j) feature maps on k-th neighbour "ring", n_r be the number of "rings" in the whole neighbourhood of x, then the method to calculate homogeneity weight ω for the whole surround is given in Eq. 3.20.

Based on these considerations, the orientation contrast of x to its k-th neighbourhood is:

$$\hat{C}_O(x, \mathcal{NH}_O(k)) = \frac{\overline{C}_O(x, \mathcal{NH}_O(k))}{\xi + \omega_k} \tag{3.19}$$

where ξ is a parameter used to prevent a zero denominator and usually set to 1.

Let m_r be the number of "rings" in a neighbourhood, and $d_{gauss}(k)$ (defined in equation (13)) be the Gaussian weighted distance of the k-th neighbourhood to x. Because of the chessboard distance, $d_{gauss}(k)$ is the same for each point within x's k-th neighbourhood. Finally, the orientation salience of x to all of its neighbours is:

$$C_O(x) = \frac{\sum\limits_k \hat{C}_O(x, \mathcal{NH}_O(k)) \cdot d_{gauss}(k)}{(\xi + \omega) \cdot m_r \cdot \sum\limits_k d_{gauss}(k)}$$

where

$$m_r = \sum_k 1 \text{ and } |\hat{C}_O(x, \mathcal{NH}_O(k))| > 0;$$

The weight ω evaluates the orientation homogeneity between x and its whole surround. It is calculated in the following way. First, given that w_{ijk} is y_i's value of the orientation (θ_j) feature maps of the k-th neighbour "ring" of x, we calculate the individual member values ($H_k(\theta_j)$) for each orientation θ_j on the individual neighbour "rings" about x and use average these as $\overline{H}(\theta_j)$:

$$H_k(\theta_j) = \sum_{y_i \in \mathcal{NH}_O(k)} w_{ijk}(\theta_j, y_i); \;\; \theta_j \in [\theta_1 \ldots \theta_\zeta]; \;\; \overline{H}(\theta_j) = \frac{1}{n_r} \sum_k H_k(\theta_j)$$

Then we globally normalize them to generate the final evaluation weight ω of orientation homogeneity:

$$\widehat{H}(\theta_j) = \sum_k \frac{\left| H_k(\theta_j) - \overline{H}(\theta_j) \right|}{MAX\left\{ H_k(\theta_j), \overline{H}(\theta_j) \right\}}; \;\; \omega = \sum_j \widehat{H}(\theta_j) \tag{3.20}$$

3.3.4.3 The Salience of a Grouping

Suppose x_i is an arbitrary component within a grouping $\Re$. Here, x_i may be either a point or a sub-grouping within $\Re$. Then the visual salience S of a grouping $\Re$ is obtained from the following formula:

$$S(\Re) = \gamma_{CI} \sum_i S_{CI}(x_i) + \gamma_O \sum_i S_O(x_i) \tag{3.21}$$

where γ_{CI}, γ_O are the weighting coefficients for the colour-intensity and orientation salience contributing to the grouping salience. $\sum_i S_O(x_i)$ is computed from the primary oriented components of grouping $\Re$ but not from the shape of $\Re$ itself. The shape distribution or boundary of a grouping may be arbitrary and may conflict with orientations of the components in the grouping. This causes some uncertainty about how to evaluate the direction of a grouping. Here the model employs a simple statistical method to deal with this problem (see [22] for other complex statistical methods involved in this field).

Suppose that x_{i_0}, ..., x_{i_j}, ..., $x_{i_{n_0}} \in \Re$ are components of a given grouping with orientation components θ_0, ..., θ_j, ..., θ_{n_0} respectively. $C_O(x_{i_j};\theta_j)$ is the orientation salience of x_{i_j} with orientation θ_j, $\hat{O}$ denotes the primary orientation on which (orientation) map the grouping $\Re$ has the maximum sum value at the current layer of the orientation pyramids. A simple method to compute $\hat{O}$ is: calculate the sum on each θ_j orientation map of all components within $\Re$ to obtain a distribution histogram of different oriented vectors (as the horizontal ordinates); then take the orientation which has the maximum value in the histogram. The formula for calculating $\sum_i S_O(x_i)$ is then:

$$\sum_i S_O(x_i) = \sum_i C_O(x_i) \; when \; \theta_j = \hat{O} \tag{3.22}$$

The above formulae for the salience computation of a grouping is a practical implementation based upon the theory discussed in Eq. 3.7.

As mentioned before, some other factors influencing salience are not considered at the moment, for example, the relative size factor between a grouping and its surrounding groupings. When the size of a target is different from the surrounding distractors but shares all other properties with these distractors, the target will "pop-out". The current computation method is inapplicable in this special case. This factor looks very simple and seems easy to implement but it is not in practice. There are a lot of problems associated with it and some are difficult to resolve. One problem is how to evaluate the homogeneity of the target's surround, especially to surrounding objects or regions.

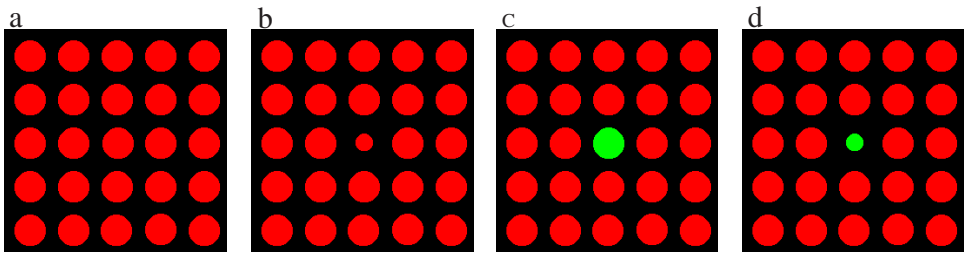

Figure 3.6: An example for salience varying with relative size between the center target and surrounding distractors

The homogeneity of a surround is affected by many factors such as shape, orientation, or colour. The shape of an object or a region may be arbitrary, so the "pop-out" by the relative size factor would depend upon the shape factor as well even if excluding other factors such as how to quantify the relationship between salience and the relative size.

Another problem is how to evaluate the degree of homogeneity and heterogeneity of the surround of a grouping, especially under the consideration of orientation. The method (shown in Eq. 3.18, 3.19 and 3.20) used here is simple and may work under many homogeneous or heterogeneous environments. For example, the homogeneity surround: all neighbours with the same orientation should be different from another homogeneity surround: some different neighbour rings have (some) different orientations but on each ring the neighbours have the same orientation. But this method is not complete especially when the surround consists of arbitrary objects. As mentioned above, an object has a shape and the shape may be arbitrary. Even if ignoring other factors such as colours, how to calculate an object orientation is not easy and this directly affects the homogeneity of a surround. The difficulty is that there is no reference which can be used to evaluate an exact order of the different homogeneity distributions of orientations. Solutions for the above problems need more evidence from other research fields such as psychophysics and neuroscience.

Figure 3.6 shows an example about the relative size factor. In Figures 3.6 (a) and (b), the red target "pops-out" in (b) when it becomes smaller. But in Figures 3.6 (c) and (d), which green target is more salient? Although (c) and (d) are the same to (a) and (b) except the target's colour, it may be that the target in (c) is more salient than the target in (d).

3.3.4.4 Grouping Saliency Mapping

Grouping saliency mapping here is a saliency pyramid in which multiple saliency maps are formed by mapping the saliency of all groupings located at different layers of feature maps into the corresponding saliency map. At any given time, this saliency pyramid is built independently at each level from the finest resolution to the coarsest one (i.e., from the lowest pyramid level to the highest pyramid lever) and dynamically varies with attention movements. At a given time, the saliency of groupings created from different feature pyramid levels (i.e., different resolution scales here) forms the pyramid-like saliency mapping.

Here, the purpose of building the saliency mapping as a pyramid is to implement object-based visual selection from coarse to fine resolutions when the selection system does not incorporate a space variant sensing mechanism for sampling a scene in a resolution-varying way. For this purpose, HOAM employs a multi-level saliency mapping and uses multiple saliency maps at different scales rather than integrating saliency maps across multiple resolutions. When employing a retina-like sensor in a scene to obtain a resolution-varying foveated image which is used for the saliency calculation, the integration of saliency across multiple resolutions will be required, because with gaze shifts over time a specific location in the scene will cross multiple resolutions and hence has different salience when viewed at different fixation locations (in different foveated images). The approach for this saliency integration will be described in Section 6.4.1.

Suppose $SaliencyMap(t)$ is a grouping-based pyramidal saliency mapping at time t, $\vec{\lambda} = [\lambda_1 \cdots \hat{\lambda} \cdots \lambda_n]$ is a vector representing for multiple pyramid levels (here it is equal to the resolution scales) from the lowest level λ_1 (the finest resolution scale) to the highest level λ_n (the coarsest resolution scale), $\overrightarrow{SM}(\vec{\lambda},t)$ is the pyramidal saliency mapping built along $\vec{\lambda}$. Let $SM(\hat{\lambda},t)$ be the saliency map at the $\hat{\lambda}$th level of the saliency pyramid, $S_{i,j}(\Re)$ be the saliency of grouping $\Re$ mapped into $SM(\hat{\lambda},t)$ at the position (i,j) at time t. Then the grouping saliency mapping $SaliencyMap(t)$ at time t can be represented as:

$$SaliencyMap(t) = \overrightarrow{SM}(\vec{\lambda},t)$$

$$\overrightarrow{SM}(\vec{\lambda},t) = \Big[SM(\lambda_1,t) \cdots SM(\hat{\lambda},t) \cdots SM(\lambda_n,t)\Big]$$

$$SM(\hat{\lambda},t) = \begin{bmatrix} \cdots & \vdots & \cdots \\ \cdots S_{i,j}(\Re) \cdots \\ \cdots & \vdots & \cdots \end{bmatrix}_{\hat{\lambda},t} \tag{3.23}$$

With the help of the above approach, attention can work on multiple saliency maps at different resolution scales to achieve object-based hierarchical selectivity from coarse to fine spatial scales. To a constant resolution scene or a uniform sampling image, normalizing saliency maps across all resolutions to build a single saliency map was used in some machine vision models to drive covert attention [65]. However, it is better that whether and how to combine saliency across resolutions are determined by current visual behaviour in a visual environment. Furthermore, it is known that human vison has resolution-varying sampling sensing so that it can not only quickly explore a scene but also flexibly acquire interesting information through "far to near" and "coarse to fine" by the limited visual processing resources. The grouping saliency mapping proposed here and in Section 6.4.1 may be better and more useful than the previous approaches that combined all scales in a resolution-uniform sampling scene.

3.3.5 Competition Pool of Attention

In this module, different groupings dynamically formed at different resolutions start to compete for attention selection from the coarsest level to the finest level by visual saliency interacting with top-down attentional biasing. The output is the dominant signal of the competitive winner(s) which is used to control the preferential processing or selectivities of visual attention. According to [23, 24, 33], the competition for visual attention can occur at multiple processing levels from low-level feature detection and representation to high-level object recognition in multiple neural systems. Also, "attention is an emergent property of many neural mechanisms working to resolve competition for visual processing and control of behaviour" [23]. The above studies provide the direct support for the integrated competition for visual attention by binding object-selection, feature-selection and space-selection. The grouping-based saliency computation and hierarchical selectivity process proposed here, therefore, is a possible approach for achieving this purpose. Hierarchical selectivity operates on the obtained grouping saliency mapping – saliency pyramid through interaction of bottom-up visual saliency from various groupings at each resolution and top-down attentional setting in the space-time context.

The outline of the top-down attentional setting logic is shown in Figure 3.7. It is

implemented as a control set of four attentional states for the current bottom-up visual input at any competitive moment:

1. *Positive priming* by which consistent bottom-up input will gain a competitive advantage;
2. *Negative priming* which is contrary to positive priming;
3. *Aimless* or *free* state in which visual attention presents a neutral state to any visual input and thus the competition for attention is completely decided by bottom-up visual saliency;
4. *Unavailability* state in which visual attention is occupied at the moment. It means no visual attention is available.

As pointed out in [46, 72], top-down priming and bottom-up visual saliency both play important biasing roles in attention capture. Top-down biasing signals affect the competition for selective attention by increasing or decreasing the baseline of neural activity. Until sufficient psychophysical findings are found to show how top-down influence directly amplifies or reduces the intrinsic salience of targets, it is feasible to take the top-down setting into the threshold of attentional competition as proposed below. If employing a competitive neural network such as a WTA (Winner-Take-All) network, a top-down setting could be implemented by installing a dynamic threshold for neuron firing but the overall computational cost for dynamic attention competition is expensive. A complex structure with an enormous number of neurons with population competition is needed.

The solution presented here is to implement attentional setting via a threshold at the decision-points in the hierarchical selectivity process. Top-down attention setting here

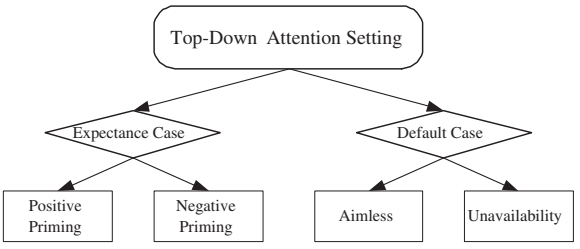

Figure 3.7: Top-down attention setting

colour flag	colour input	orientation flag	orientation input	"view details" flag

Table 3.1: Top-down attention setting to the basic features

plays two roles: one is top-down biasing for globally and locally attentional competition; another is an intention request of whether to "view details" of a grouping (e.g. its sub-groupings) when attention is deployed at a grouping. However, top-down priming for special objects or groupings is very complicated since intricate object recognition from higher level processing is at least required. At present, top-down biasing here aims to act only on the level of basic features which are the colour, intensity, and orientation feature pyramids.

The top-down signals (Table 3.1) include two flags for colour (including intensity) and orientation top-down biasing and one flag for "view details". Each flag encodes the states of its correspondingly top-down signal. For colour and orientation flags, "00" is the default case in which all groupings compete for visual attention in the pure bottom-up way; "01" encodes positive priming in which all groupings with the positively primed feature preferentially compete for attention and at the same time other competitors are suppressed; "10" encodes negative priming which is the inverse to positive priming; "11" is the unavailable state in which all groupings having these features are prevented from attracting visual attention. For the "view details" flag, "0" signals "continue" to explore details of a grouping (i.e. its sub-groupings if they exist at the current resolution or the finer resolution) and "1" means "shift" attention from the current winner to the next potential winning grouping. The next winner will be generated from the unattended groupings at the same resolution as the current winner if these groupings exist, otherwise from the unattended groupings that lie at the same coarser resolution as the parent grouping of the current winner (see hierarchical selectivity below). This process links from the "lineal chain" to the "collateral chain".

Hierarchical selectivity operates on the interaction between grouping salience and the top-down attentional setting at any competitive moment. The competition for visual attention occurs first among the coarsest groupings (existing at the coarsest resolution) by global competition. Through a WTA (Winner-Take-All) mechanism, visual attention is firstly deployed to the winning competitor. Then, a top-down or goal-driven (request) control of whether "continuing" to view the details within the current group-

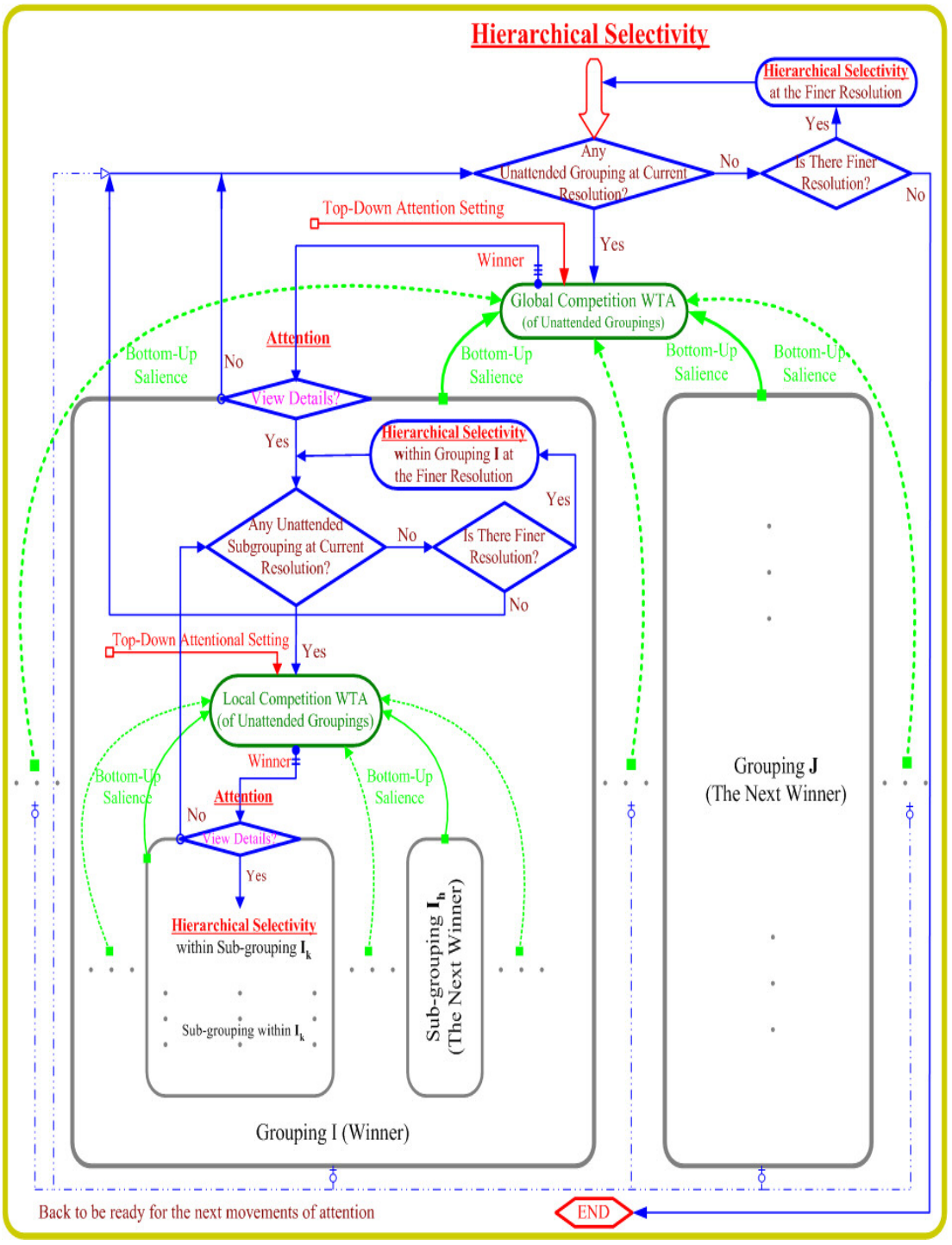

Figure 3.8: Diagram of hierarchical selectivity: see text for detailed explanation.

ing or "shifting" attention out of this grouping takes place.

If switching attention, the next winning competitor gains visual attention with the aid of an "inhibition of return" mechanism which prohibits attention from instantly returning to a previously attended winner. The priority order for generating the next potential winner is:

1. The most salient unattended grouping that is a sibling of the current attended grouping. This winning grouping has the same parent as the current attended grouping and both lie at the same resolution.

2. The most salient unattended grouping that is a sibling of the parent of the current attended grouping, if the above winner can not be obtained.

3. The backtracking continues if the above is not satisfied.

Temporary inhibitions to the attended groupings can be used to implement inhibition of return (presented in Chapter 6). More elaborate implementations may introduce dynamic time control into different winners so that some previously attended groupings can be visited again. But the work here assumes that each winner is attended only once.

If continuing to check the current attended grouping, the competition for attention based on bidirectionally bottom-up and top-down interaction by local competition is triggered firstly among the sub-groupings that exist at the current resolution and then among the sub-groupings that exist at the next finer resolution. This indicates that the sub-groupings at the finer resolution do not gain attention until their siblings at the coarser resolution are attended. By the force of WTA, the most salient sub-grouping wins visual attention.

After attention has been directed to the winning grouping/sub-grouping, the same (top-down) goal-driven method is used to determine whether or not to "continue" to look into the details within this grouping/sub-grouping. If not, another attention "shift" takes place. If continuing to examine the particulars of this grouping/sub-grouping, another local competition triggers. When "continuing" to check an attended grouping/sub-grouping is requested, if there is no sub-grouping existing at the current or a finer resolution, hierarchical selectivity goes back to the parent of the current attended grouping. At this moment, the same "continuing/shift" attention occurs. This "continuing/shift" recursive procedure continues until the desired goal is reached or all groupings in a scene are attended.

As mentioned before, the grouping saliency computation is independent of how to segment the groupings in a scene. The mechanism for hierarchical selectivity is also

independent of what/how a segmentation is used at multiple resolutions or a single resolution for a scene. The choice of segmentation or grouping method is not included in these two mechanisms. Hierarchical selectivity runs on a given segmented scene and is driven by both the top-down attentional setting and the current distribution of the given segmentation and the corresponding salience. Switching attention between groupings/sub-groupings (and between the coarse and fine resolutions if multiple resolutions are used) is then controlled. A diagram summarizing the recursive procedure for hierarchical selectivity is given in Figure 3.8. Its algorithmic description is given in Table 3.2.

Two goals can be achieved by taking advantage of hierarchical selectivity. One is that attention shifting from a grouping to another and from groupings/sub-groupings to sub-groupings/groupings can be easily carried out. Another is that the model may simulate the behaviour of humans observing something from far to near and from coarse to fine. Meanwhile, it also easily operates at a single resolution level. In addition, a declaration we made here is that the top-down attentional setting in hierarchical selectivity is not completely implemented although its possible approach is given in the algorithm. Except for "colour-flag=00", "orientation-flag=00" and "view details flag", other cases will be realized in the future.

Support for this approach to hierarchical selectivity has been found in recent psychophysical research on object-based visual attention. It has been shown that features or parts of a single object or grouping can gain an object-based attention advantage in comparison with those that are separated from different objects or groupings. Also, visual attention can occur at different levels of a structured hierarchy of objects at multiple spatial scales. At each level all elements or features coded as properties of the same part or the whole of an object are facilitated in tandem (see [6] for a review of these viewpoints and detailed findings).

3.3.6 Perceptual Grouping

It has been suggested [6] that grouping processes and perceptual organization play an integral role in object-based attention. Features that are grouped together compete against other feature groupings and obtain faster processing than features that do not belong together. Perceptual grouping is a complex combinatorial problem which involves a lot of influence factors including top-down interference in many conditions. These factors work together to affect how groupings are segmented, such as spatial

1. competition begins among the coarsest groupings at the coarsest resolution;
2. if (no unattended grouping exists at the current resolution) goto step 10;
3. unattended groupings at the current resolution are initialised to compete for attention based on their salience and top-down attentional setting;
4. check the colour-flag and orientation-flag and apply corresponding top-down processing to modify the active states of the groupings (details are not implemented here);
5. all (modified) groupings compete for attention;
6. attention is directed to the winner (the most salient grouping) by the WTA rule; set "inhibition of return" to the current attended winner;
7. if (the desired goal is reached) goto step 12;
8. if ("view details" flag=1) (i.e. don't view details and shift the current attention)
   { set "inhibition" to all sub-groupings of the current attended winner; }
   if (the current attended winner has unattended siblings at the current resolution)
   { competition starts between these siblings; goto step 2 and replace the grouping(s) by these siblings; }
   else goto step 11;
9. if ("view details" flag=0) (i.e. continue to view the details of the current attended winner)
   if (the current attended winner has no sub-grouping at the current resolution)
   goto step 10;
   else { competition starts between the winner's sub-groupings at the current resolution; goto step 2 and replace the grouping(s) by the winner's sub-groupings; }
10. if ((a finer resolution exists) and (unattended groupings/sub-groupings exist at the finer resolution))
    {competition starts on groupings/sub-groupings at the finer resolution; goto step 2;}
11. if (the current resolution is not the coarsest resolution)
    { go back to the parent of the current attended winner and goto step 2; }
12. stop.

Table 3.2: The algorithmic description of hierarchical selectivity

proximity, similarity, common fate, shared properties, and even experience and learning [105, p.257-309]. In many cases, the rules for segmentation and interpretations of groupings are associated with visual tasks and experience. Nevertheless, a study of this field is out of the current scope of our research. The groupings used in this chapter are produced by manual preprocessing based on Gestalt principles and heuristic knowledge, to provide the basis for experiments with our attentional model.

The principles of grouping used are some common rules: proximity, closure, continuity, common fate, familiarity, and shared properties. A visual grouping is defined as an effective hierarchical structure formed by all components according to these principles. For example, objects that share a common colour or orientation and are separated from their surrounding which does not share this colour or orientation may be organized as a grouping. Objects belonging to a large group or share the same spatial location may be segmented into a multi-level structured grouping. In Figure 4.11, the white stripes in the road are grouped into three groupings by their familiarity. The four cars are organized as a grouping by their common fate. Two people are grouped together by their proximity.

In fact, the "grouping" addressed here is the "perceptual unit" which serves as the potential unit of attention deployment. For object-based attention, it is the "proto-object" produced by various segmentation processes rather than the conceptual or recognizable "object" we commonly experience in the real world. "Evidence suggests that 'object-based' attention and 'group-based' attention may reflect the operation of the same underlying attentional circuits" [122]. One general criticism of object-based attention is the question of whether objects are recognized before or after the selectivity by attention, or whether visual segmentation processes occurs with attention or without attention. This is also the traditional story in terms of "early selection" and "later selection" or the degree of preattentive processing in the visual systems. The issues stressed here may lead to a better understanding of the grouping mechanism. A further discussion on these issues can be found in [27, 122].

Following the above theoretic description, the next chapter will use both synthetic and natural images to demonstrate the performance of the proposed object-based attention model HOAM. The related conclusion and discussion are also reported.

Chapter 4

Performance on Synthetic images and Natural Scenes

The program used to implement the model is C++ and ran the experiments reported in this thesis on a Sun Ultra 4 workstation. In general, it took about two to ten minutes to complete an experiment depending on different image sizes. The next section examines the performance of the Hierarchical Object-based Attention Model (HOAM) in both simulated psychophysical displays and real-world natural scenes. The experimental results show that the behaviour of our model concurs with the main findings found in psychophysical research on visual attention and the successful performance in hierarchical selectivity. Finally, the last section summarises the overall results made in this chapter and suggests some useful future work.

4.1 Examination and Discussion

For the evaluation of proposed object-based attention model, a number of experiments based on synthetic and natural scenes were run.

4.1.1 Performance in Synthetic Images

The goal of the experiments in this section is to demonstrate the performance of our model concurring with the main findings found in psychophysical research on human visual attention. The experiments below are designed for this purpose.

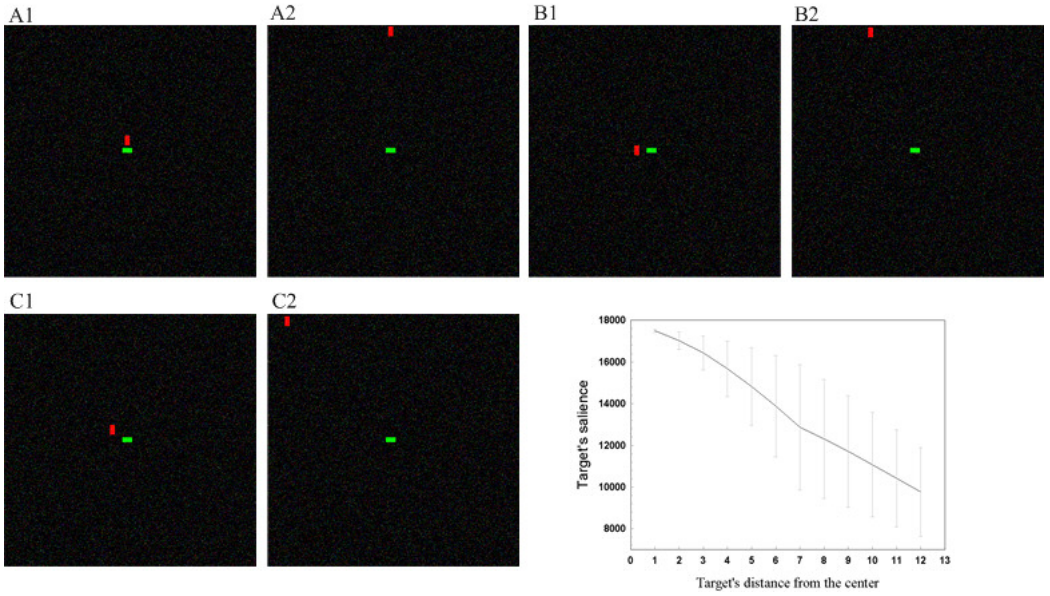

Figure 4.1: Object position experiments. The target is the red vertical bar and the green horizontal bar is the reference object located in the center. Two of the displays used for each sub-experiment (A, B, and C) are shown respectively. Solid lines: relationship between object's position and its salience. Error bars: one standard deviation (15 runs at each distance for three directions).

4.1.1.1 Object Position Influence

Many visual search studies in psychophysics have suggested the effects of eccentricity on visual search, that is, visual search is less effective when the target location is farther from the fixation center [148]. We have produced synthetic images to test the performance of the model on this point. The first and last series of the test images are shown in Figure 4.1. All the images are 512×512 colour images with added 30% random colour noise (30% random pixels in the image are affected by adding random colour pixels to this image. The noise filter alters each affected pixel based on its current colour and the colours of any adjacent pixels. The similar noise filter will be used in all other experiments if there is no special explanation for adding noise effect). All stimuli here are the same size and the target is always the red vertical bar. In all experiments, the fixation center is assumed at the center of the images. The target initially lies closest to the referenced object located at the center in the image, and then moves away gradually (the reference bar is used to conveniently calculate the target's salience and indicate the initial locus of attention). If we take each bar as a grouping in the display (like a pixel in the image), according to our Gaussian weighting definition

(shown in Eq. 3.13), bars located at the same radius have the same distance effect. For example, the bars in the images A1, B1 and C1 in Figure 4.1 lie at radius 1 and the bars in the images A2, B2 and C2 have radius 12. In this way, twelve series of images (36 images in total) were produced. Each series includes 3 images in which the red bars have a certain position relative to the green bar. The target's salience is calculated by considering the reference bar and all other pixels. Both colour and orientation contrasts are included.

This experiment examines the relationship between searching efficiency and target position variation. This aim can be achieved by probing the target's salience along its position trajectory without top-down attentional priming. One therefore does not need to calculate the target's salience on all pyramid layers (i.e. multiple resolutions) and it is harmless to remove the effects of top-down attentional priming. For a demonstration, it is sufficient to compute the target's salience on the lowest layer of the pyramids and set up top-down attentional setting to the free-state by default. (This kind of consideration runs through the following synthetic experiments by default.)

Results from the experiments are presented in Figure 4.1. Clearly, as the target's distance from the center of the display increases, its salience also declines in the experiments so that visual search becomes less effective. It can be seen that closer objects affect each other more when competing for visual attention than farther objects. The bigger error bars with farther distance indicate the greater differences between the salience of targets in the experiments A, B and C. The reason is that the target has more and more different neighbours when it lies at the same radius but different positions in the display. However, whether the influence is facilitated or not depends on the homogeneity or heterogeneity of the actual distribution between the object and its surroundings. More discussion on this issue is given in the following experiments.

However, there are some limitations to the object-based attention model HOAM in such experiments: if the target and the reference bar share all of the absolute features such as the same red colour and the same orientation (here, they are conjunctions of feature colours and orientations) and are formed in a line with only one pixel distance, then the model can not produce performance similar to this experiment for position. The reason is that the saliency computation used in the model is based on center-surround contrasts and the eccentricity factor is integrated into the saliency computation of features within groupings rather than as an independent computation.

4.1.1.2 Neighbourhood Influence on Visual Search

Many psychophysical studies of visual attention (especially on object-based attention) have suggested that visual search is greatly affected by the attribute distribution and interaction between target and its surroundings (see [6, 107, 148] for a detailed explanation). These effects are clearly observed in experiments on testing similarity or shared feature dimensions between target and non-target and on homogeneity or heterogeneity of the target's surround. When the distractors surrounding the target are more homogeneous to each other and share less features with the target, search becomes more efficient or accelerated. Perceptual grouping also plays an important role, by which distractors are grouped by type so stronger grouping strength leads to easier pop-out [29, 79, 85].

Three kinds of experiments are designed to test the model performance. The experiments probe the salience variation of the target in response to the surrounding changing without top-down attentional priming. It is also not necessary to calculate the target's salience on all resolution levels. For a demonstration, it is sufficient to compute the target's salience at the finest resolution and set the top-down attentional setting to the free-state by default. (This kind of consideration runs through the following synthetic experiments by default.)

Experiment 1: The scaling effect of uniform neighbours

The experimental method is that the target is located at one place and kept fixed. Then more and more homogeneous neighbours, which have at least one feature different to the target, are added. The goal of this test is to prove that when the number of such homogeneous neighbours increases (i.e., the facilitated strength of the neighbours is stronger), the target's salience increases so that the target's pop-out becomes easier.

Two series of sub-experiments are produced to examine the model performance. In experiment A and B, the created images are all of 256×256 and the target is always a red bar located at the center of the displays. Green horizontal bars are gradually added in the neighbourhood of the target and kept homogeneous. Compared with experiment A, the target in experiment B is vertical. So, the target is different from its neighbours by only one feature of colour in experiment A, two features of colour and orientation in experiment B. Features considered in the computation of the target's salience are colour and orientation. Both distractors and background take part in the salience computation of the target. That is, the target's salience is derived from the contrast not only between

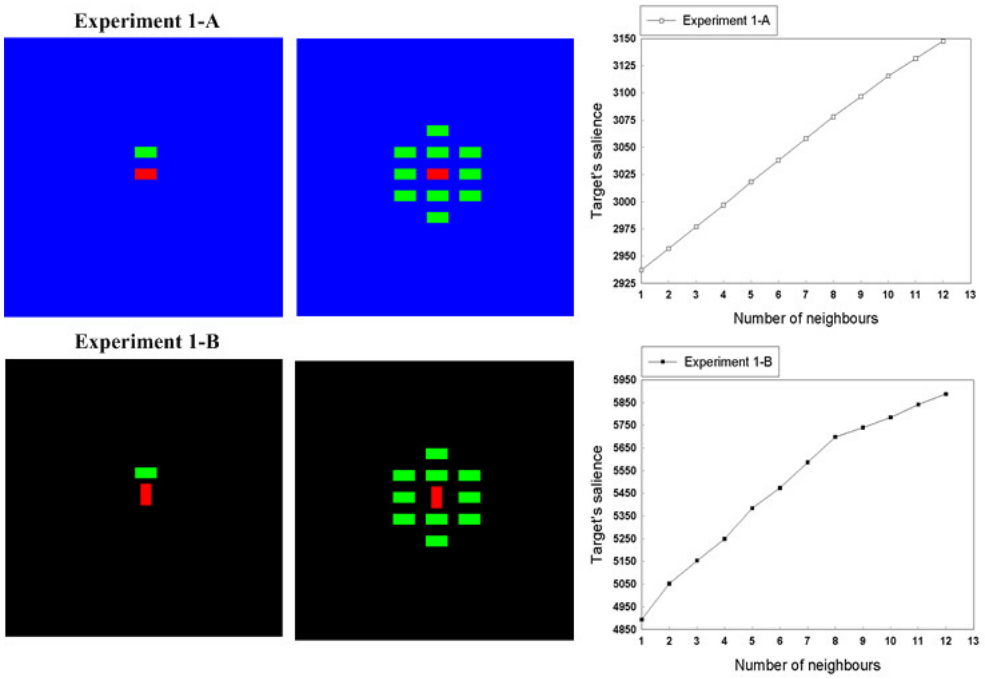

Figure 4.2: The performance of the model in experiments varying the scaling effect of uniform neighbours. Left and middle columns: two of the displays used in each sub-experiment. The related results of each experiment are shown in the right graphs respectively.

the target and distractors but also between the target and background. Figure 4.2 shows several created images and the results of the target's salience in these two experiments.

Discussion: The results from experiments A and B clearly show increasing target salience with increasing homogeneous neighbours (greater strength of neighbourhood). This is consistent with the findings from psychophysical experiments. Furthermore, the curve in experiment B ascends faster than that in experiment A (notice the different scales of Y-axes in experiments A and B). It is suggested that uniform neighbours sharing fewer features with the target make the target more salient and hence attract more visual attention. We also did another experiment based on experiment A (not presented) in which we adjusted the target's size. When the target became smaller its salience became smaller. But when the target became smaller and shared the same colour with the distractors, the results became unpredictable because of the relative size factor. As we already discussed in Section 3.3.4.3, the model will fail to perform for this special case.

Experiment 2: The effect of coherence in the target's neighbourhood

This experiment investigates the salience of the target in an originally homogeneous surrounding by gradually changing one attribute of more and more neighbours to another (colour or orientation) but keeping them homogeneous. We produced two series of test images with size 256×256 for two sub-experiments. In the first sub-experiment more and more items surrounding the target change colour to be the same as that of the target. The target's salience comes from its comparison with all other circles and green background. In the second sub-experiment, the neighbour items become orthogonal to the target one by one. In this sub-experiment, the salience of the target is derived from its comparison with all other red bars and the black background. To remove the effect of distance varying when a horizontal bar is rotated, the computation for distance factor is designed as: all red bars within the same neighbourhood have the same distance whatever their orientations are. That is, when a horizontal red bar is rotated to vertical, its distance remains the same as before. Several images and the results of these experiments are given in Figure 4.3.

Discussion: The results shows that the target's salience becomes weaker as more neighbours share the same colour as the target in experiment 2-A, but stronger as more neighbours turn orthogonally to the target in experiment 2-B. The reason is that in experiment 2-A the strength of grouping based on the colour green within the target's homogeneous neighbourhood became weaker while the strength of grouping based on colour red is stronger. In experiment 2-B, although the neighbours form two types of groupings, the new continuously growing grouping did not affect the neighbourhood homogeneity but enhanced the contrast to the target. In fact, both experiments have the same nature but reflect different aspects of the effect of the target's neighbourhood. The result of experiments 1 and 2, as pointed out in [30, 62] and other research on object-based visual attention, have shown that stronger grouping distractors and greater differences between the target and distractors enable the target to be sought more efficiently. In other words, stronger contrast between the target and its neighbourhood makes the target more salient to capture visual attention in the bottom-up competition.

Experiment 3: Effect of the target neighbourhood heterogeneity

This experiment examines the performance of the model in heterogeneous circumstances. In theory, the target should be less salient with a more disorderly distribution of the neighbourhood. The method used here is similar to the two previous experi-

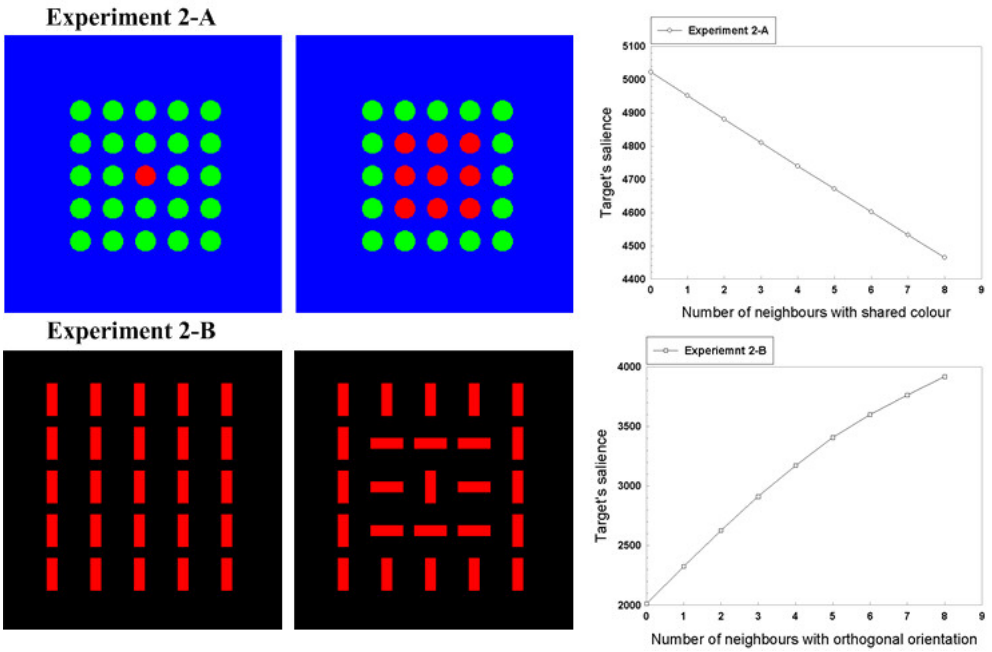

Figure 4.3: Model performance when varying attributes of target's neighbours in a homogeneous environment. 2-A: the target is a red circle located at the center of the display and the neighbours change to the colour of the target. 2-B: the target is the vertical red bar located at the center of the display and its neighbours change to the orientation orthogonal to the target. Left column: first test display. Middle column: 8th test display.

ments. The red vertical target was initially located at the center of a homogeneous surrounding in which the same colour bars are orthogonal to the target. After that, we gradually varied the neighbours' orientations to create a series of more and more heterogeneous displays. One experiment is shown in Figure 4.4. All displays have added 30% random colour noise. The target's salience is computed by colour and orientation of the target contrasting with both of the distractors and background. Although we do not give results from all the experiments, the overall experimental results are similar to that in Figure 4.4.

Discussion: The results shown in the bottom diagram in Figure 4.4 indicate that the target's salience decreases with the growing heterogeneity of its surroundings. This means that the efficiency of visual search becomes worse and worse. Notice that the downtrend of salience is much steeper in the first four steps and tends to a mild decline afterwards. The saturated tendency effect is not surprising but expected. The Gabor filter for orientation extraction used here is sensitive to four orientations of 0^0, 45^0, 90^0, and 135^0. When the number of disorderly orientations exceeds four direc-

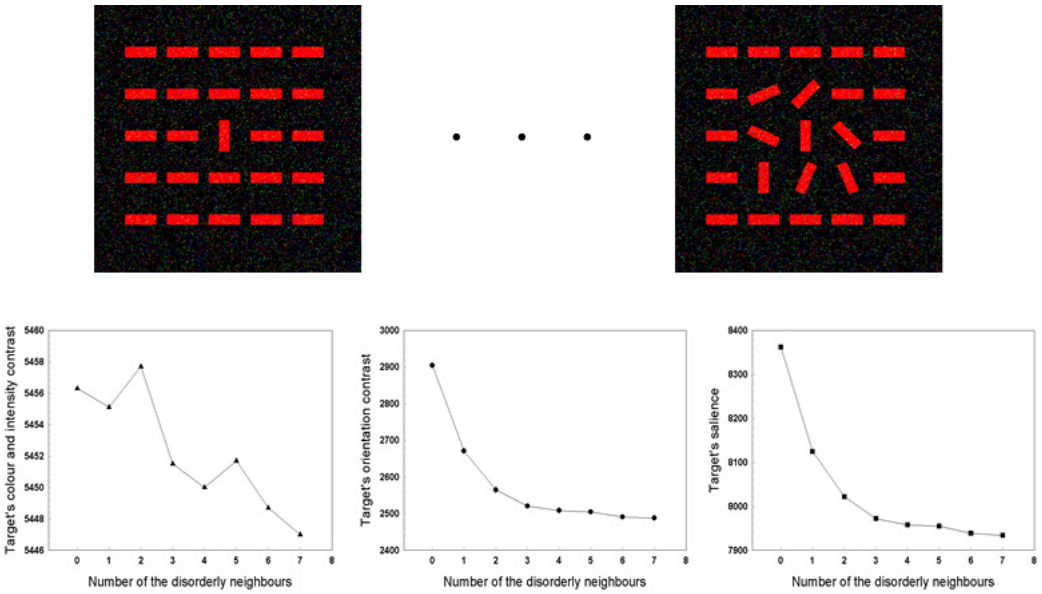

Figure 4.4: Model performance in an oriented heterogeneous environment. The target is the red vertical bar located at the center of the display. Its neighbours become more and more heterogeneous by gradually varying their orientations from the target and each other. Two members of the sequential displays are shown here. Little colour-intensity contrast effect and great orientation contrast effect shown in the two graphs at bottom left suggest that the target's salience is mainly affected by the orientation disorder factor.

tions, the result is an almost saturated weight ω in Eq. 3.20 (see Section 3.3.4.2) because ω is limited by the maximum different orientations (4 here). This ω is used to evaluate the orientation disorder within an object's neighbourhood. Another observed phenomenon from the graphs in Figure 4.4 is that the main contributor to regularly reduce the target's salience is the orientation disorder factor rather than colour effect. The explanation for this effect is that the distractors always shared the same colour with the target and the varying position of each pixel within each distractor grouping in this experiment produced only a tiny effect in the colour contrast between the target and the distractor, so the overall trend of the target's salience is hardly affected by the colour of the surrounding features.

4.1.1.3 Intensity Varying Influence

This experiment will examine behaviour of the model performance with varying intensity. Human attention research suggested that visual search becomes easier "pop out" when the contrast between the target and its surround increases [106, 148]. For evaluation convenience, a series of black-white displays are designed (Figure 4.5). In these displays, the target is the circle located at the center with a black background. The target intensity gradually changes from 255 to 0. Three levels (0%, 10%, and 30%) of random (0-255) noise are added to each display for three sub-experiments respectively. Salience of the target is calculated by comparing the target with the background for all displays. Figure 4.5 shows the experiment results.

The results show that the target salience goes up stably with increasing intensity contrast between the target and the background under non-noise and noise environments. When the intensity contrast of the target with its background is reasonable, the target salience generally declines as noise rises. We also explored similar cases using other displays with different strength (50%, 75%, and 90%) noise. The results are overall similar but the intensity range of the target in this case varies with the difference between the relative size of the target and its background.

4.1.1.4 Oriented Direction Influence

We have known that increasing orientation contrast produces a nonlinearly enhanced effect on visual salience of pop-out [102]. In Section 3.3.4.2, a sinusoid function was used to evaluate the effect of different directions of feature orientation on grouping salience in a given condition to reach this findings. Here we give a very simple ex-

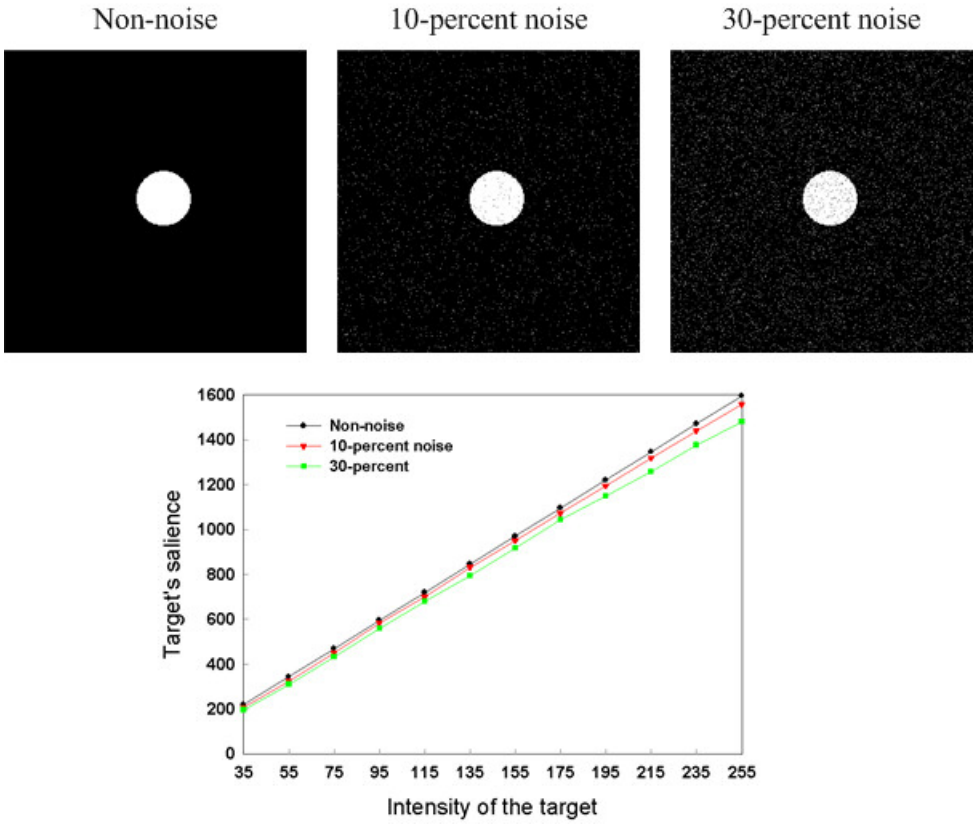

Figure 4.5: Testing the effect of intensity contrast on visual salience. In the displays the target is always the circle object at the center. One member of the displays is shown for each sub-experiment with different strength noise.

periment (Figure 4.6) to test the model's performance on this effect. All displays used here are of size 256×256. All items in each display share the same size and colour and the target is always located at the center of the display. The target changed its original direction (initially 0^0) to 22.5^0, 45^0, 67.5^0, and 90^0 respectively while keeping its size and colour constant. Because of the symmetry effect of the orientation filter used here, it is not necessary to test the other three orientations. For the test purpose, we adjusted the Gabor filter to be sensitive to 8 orientations. The contributors to the target's salience still include colour, orientation, distractors and background. Figure 4.6 shows the target's salience increasing stably with increased orientation contrast between the target and its surroundings. Meanwhile, the salience curve grows up quickly from zero degree to around 30 degree and climbs more slowly from around 45 degree to the maximum 90 degree orientation difference between the target and its homogeneous neighbourhood.

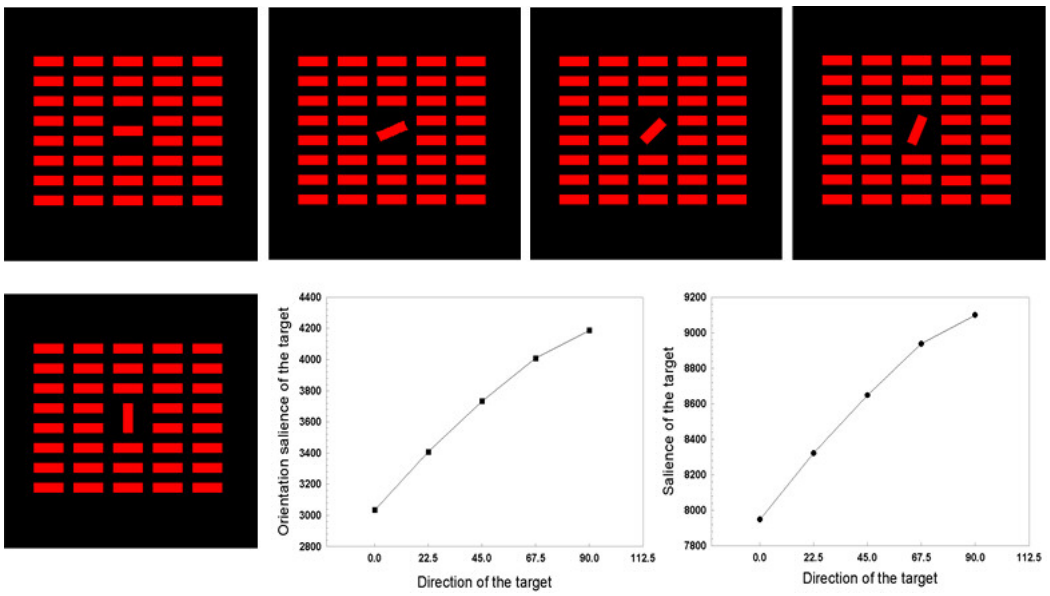

Figure 4.6: Testing the effect of orientation contrast on visual salience. In the displays the target is always the red bar at the center.

4.1.1.5 Grouping Influence and Related Hierarchical Selection

Because the presented model uses hierarchical grouping-based saliency mapping and competition for integrated object and space-based attention, the group advantage for visual search [30] becomes an inherent property of the model. Here an experimental example is used to show how the model works for hierarchical selectivity. Figure 4.7 shows a display in which the target is the only vertical red bar and no one of the bars has exactly the same colour as another bar. Three bars have the same orientation and others have different orientations. If we do not use any grouping rule, each bar may form a single grouping by itself. Then we obtain 36 single groupings. If segmenting the display by shared orientation, the only structured grouping is formed by the 3 vertical bars, which includes the red target (forms one sub-grouping) and other two vertical green bars (forms another two-level sub-grouping). In this way, 34 top groupings (38 in total) can be obtained: a structured three-level grouping (contains 4 sub-groupings) and 33 single groupings formed by other distractors respectively. The resulting saliency maps and attention sequences for these two segmentations are given in Figure 4.7. The background, colours, and orientations are considered in the salience computation. The top-down attentional setting is set to the free state, so this is pure bottom-up attention competition.

The results show different orders of paying attention to the targets. The three ver-

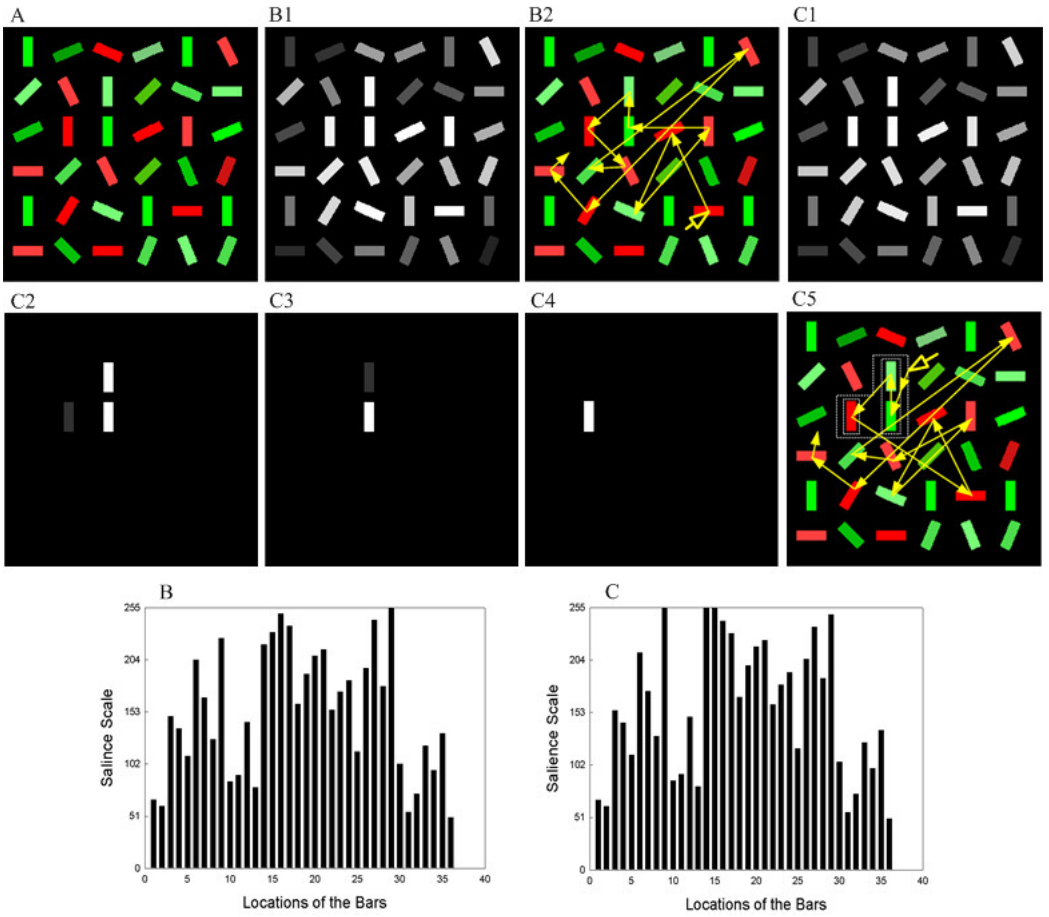

Figure 4.7: An example for structured groups and hierarchical selection. In the display the target is the vertical red bar at the third row and the second column. B1: saliency map (in shades of grey) in the case of no grouping. B2: partial attention sequence of most salient bars for B1. C1: saliency map in the case of grouping. C2, C3, C4: saliency maps of the grouped bars. C5: partial attention sequence of most salient bars for C1. B, C: salience histograms for B1 and C1 respectively. Note target is attended after 7 shifts in B2 but only 3 shifts in C5.

tical bars including the target (see Figure 4.7 (C1), (C2), (C3), (C4), and (C5)) has an advantage by group in attracting attention much more quickly than the non-grouped. The competition starts among different groupings in the display. The structured grouping of 3 vertical bars is the most salient compared to others and obtains attention firstly. Then the competition occurs within this grouping between the target and another sub-grouping formed by the two vertical but different colour bars. Attention is directed to the sub-groupings according to their salience orders when we do not consider top-down attentional priming. The target is attended after the two-level sub-grouping is attended. This grouping advantage for attentional competition has been confirmed by psychophysical research on object-based attention [6, 122].

4.1.2 Performance in Natural Scenes

We showed the examination of the proposed model (HOAM) on visual (covert) attention behaviour by using some artificial images and successfully demonstrated the results compatible with related findings in psychophysical research of visual attention. To investigate the model in complex natural scenes, colour outdoor photographs taken with a digital camera are used here. The implementations for both of "from coarse to fine" and "from far to near" human attention simulations in these real-world images are described in detail.

4.1.2.1 Hierarchical Selectivity

As suggested in [122], "there may be a hierarchy of units of attention, ranging from intra-object surfaces and parts to multi-object surfaces and perceptual groups". Hierarchical selectivity is a novel mechanism designed for shifting attention from a grouping to another one or from a parent grouping to its sub-groupings as well as implementing attention focusing from far to near or from coarse to fine. It can work under both multiple (or variable) resolutions and single resolution environments. Resolutions can be either scaled by a pyramid decomposition scheme or by a digital camera. Here an outdoor scene is used to demonstrate the behaviour of hierarchical selectivity. In Figure 4.8, the same outdoor scene is photographed from far and near distances respectively so that two coarse (64×64) and fine (512×512) resolution photographs are obtained. In the scene, there are two groupings: a simple shack in the hill and a small boat including five people and a red box within this boat on a lake. The people, red box, and the boat itself constitute seven sub-groupings respectively for this structured grouping.

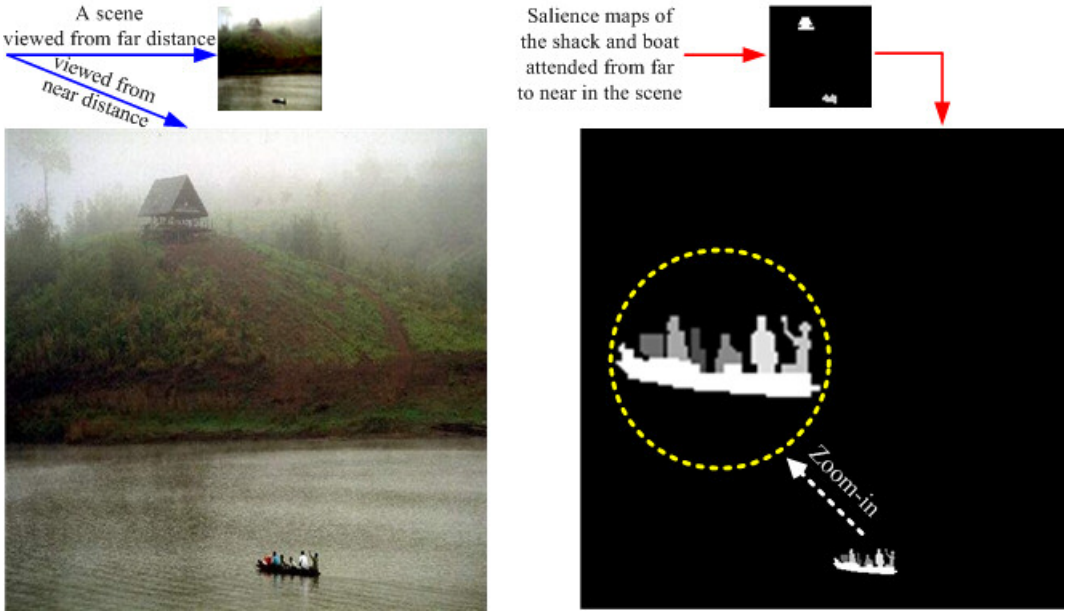

Figure 4.8: An outdoor scene photographed from far and near distance respectively. The obtained images shown here are the same scene but different resolutions. The saliency maps are shown too and the grey scales indicate the different saliency of the groupings.

The $1/\rho$ parameter for Gaussian weighting (Eq. 3.13) is set to 25% and the Gabor filter is sensitive to 4 orientations $[0^0, 45^0, 90^0, 135^0]$.

The model works with these two images, using the coarse and fine images as different resolution levels. For this purpose, only feature (colour, intensity, and orientation) maps at the lowest level of the pyramids are created for each image. (Multi-level pyramids used to simulate attending a complicated natural photograph from far to near and coarse to fine are also implemented in this chapter. See the next sections for details.) Competition for attention starts in the coarse or far image (Figure 4.9). Using hierarchical selectivity, attention is firstly deployed to the winner (here the shack) and suppresses other competitors. Then attention shifts to the fine image for further checking this winner if answering "yes" to the "view details" flag. If the answer is "no", the model will check if there is(are) any other grouping(s) existing at this image.

When attention is shifting, an "inhibition of return" is set for this attended grouping. Because the shack has no sub-groupings, attention switches again to the coarse image and checks if there exists any next winner. Thus the boat grouping obtains attention. In the same way as attending the shack, answering "yes" to the "view details" flag attention shifts to its sub-groupings in the fine image. At this moment the competition for attention triggers among the seven sub-groupings. Attention is deployed

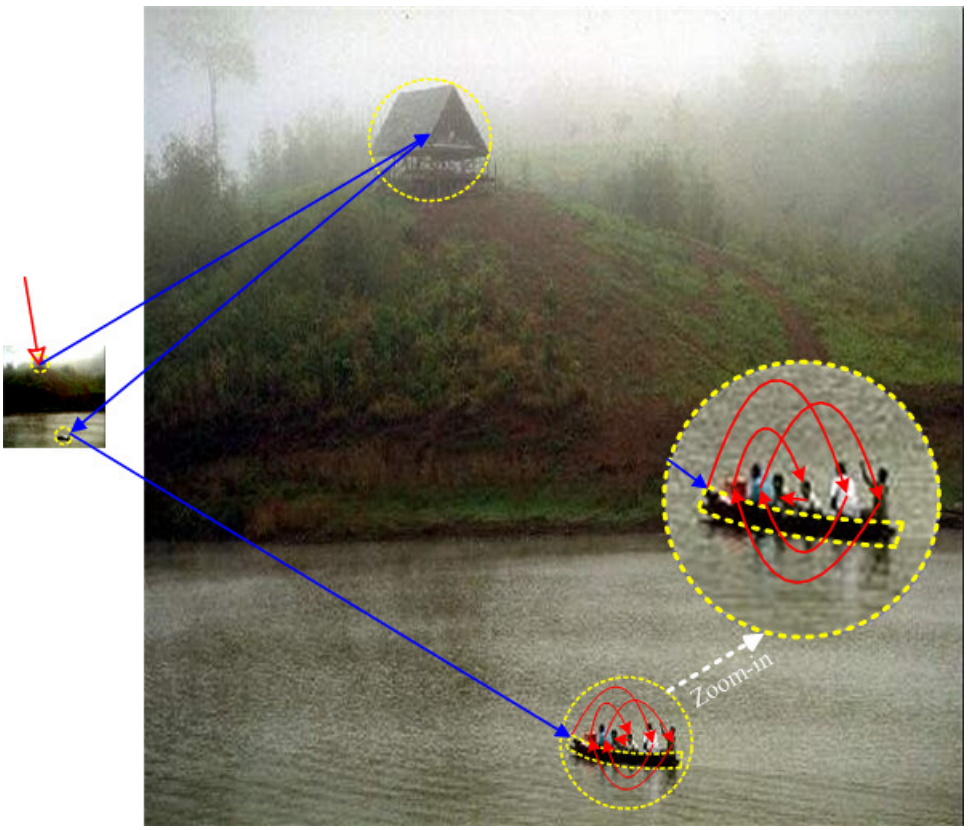

Figure 4.9: The attention movements implemented for the outdoor scene: blue arrows indicate attentional movements between resolutions and red arrows denote attention shifts at fine resolution.

to these sub-groupings by hierarchical selectivity. The saliency maps computed for these groupings are shown in Figure 4.8 and the sequence of attention deployments is shown in Figure 4.9. The attention deployment trajectory shown in Figure 4.9 reveals reasonable movements for this natural scene.

4.1.2.2 Hierarchical Selectivity From Coarse To Fine

The image presented in Figure 4.10 has 512×512 pixels and contains many structured objects and groupings. The pyramids in the model used here have three layers, ranging from resolutions 128×128 to 512×512. The Gabor filter was set to be sensitive to 4 orientations $[0^0, 45^0, 90^0, 135^0]$. The $1/\rho$ parameter for Gaussian weighting was set to 50%.

The model firstly extracted colours, intensity, and orientations from the photo and constructed altogether 9 three-layer pyramids: one intensity pyramid (Figure 4.12), four colour pyramids (Figure 4.13), and four orientation pyramids (Figure 4.14). Eleven

meaningful groupings of objects were created manually by preprocessing according to Gestalt grouping rules (see Section 3.3.6). Figure 4.11 shows the identifiers of the different groupings in this image. The numerals pointed to by each white arrow denotes the identifier of each grouping at multiple resolutions. The groupings which have the same prefix identifier belong to the same parent grouping. The depth of each grouping is the index of its array mark. For example, identifier 1-1 indicates that this is the first sub-grouping of grouping No. 1. Identifier 1-1-2 denotes it is the second sub-grouping of grouping No. 1-1. Groupings No. 1-1-1 and No. 1-1-2 have the same parent grouping No. 1-1. The black circles or ellipses are used to conveniently distinguish different groupings (object(s) in the circles) and not the grouping boundaries. When viewing these groupings at different resolutions, some groupings/sub-groupings will disappear at the lower resolution. The hierarchy of groupings is shown in Figure 4.11 and Figure 4.23 which is discussed later.

The top-down attention setting was always set to the free state in this test. The decision-points during hierarchical selectivity to drive whether or not viewing the details within a grouping were always answered "Yes". Although this may make hierarchical selection look like an exhaustive exploration, the general performance of the model can be inspected in detail and completely (see Section 4.4.3 for an alternative implementation in this view). As we discussed in Section 3.3.5, the control for recognizing which object is significant is very intricate and needs higher visual processing related to the current visual tasks (also see the following discussion about the small white stripes in this scene). Future work will refine this complicated control. In the more normal scenes, top-down priming proposed for the "view details" flag will control choice to produce more interesting behaviour.

Here, the competition for visual attention was firstly triggered at the coarsest resolution, namely the highest layer of the pyramids. During the attentional movements, shifting into the higher resolution (lower layers of pyramids) or switching to the lower resolution (higher layers of pyramids) dynamically changed depending on the natural structure of the current grouping being attended and its surroundings. When a structured parent grouping is attended at high resolution, some/all of its sub-groupings will be attended next at this current resolution if these sub-groupings appear at the same resolution, or at the lower resolution if some/all of its sub-groupings do not appear at the current resolution. In this procedure, some sub-groupings within a parent grouping, such as some small white stripes in the road, may have not much significance and may not necessarily be attended. This further top-down control for shifting attention

Figure 4.10: An outdoor photograph.

will need additional theory to incorporate the measure of object similarity, subject's experience and the current visual task, etc. and is not implemented here. The results of the model performing on all resolution levels are shown in Figures 4.15, 4.16, and 4.17.

At each attentional deployment, the results show the entire or unitary salience of the grouping which is currently being attended. When the related groupings are ready to compete for visual attention the degrees of their individual salience (in shades of grey) are presented in comparison with all other competitors. The brighter a grouping is, the more salient it is. It is worth noting that no mosaic appearance [1] is seen in the results

[1] In a location-based saliency map if all pixels within a grouping have different saliency values, they may appear different grey blocks in the map. This is not seen here.

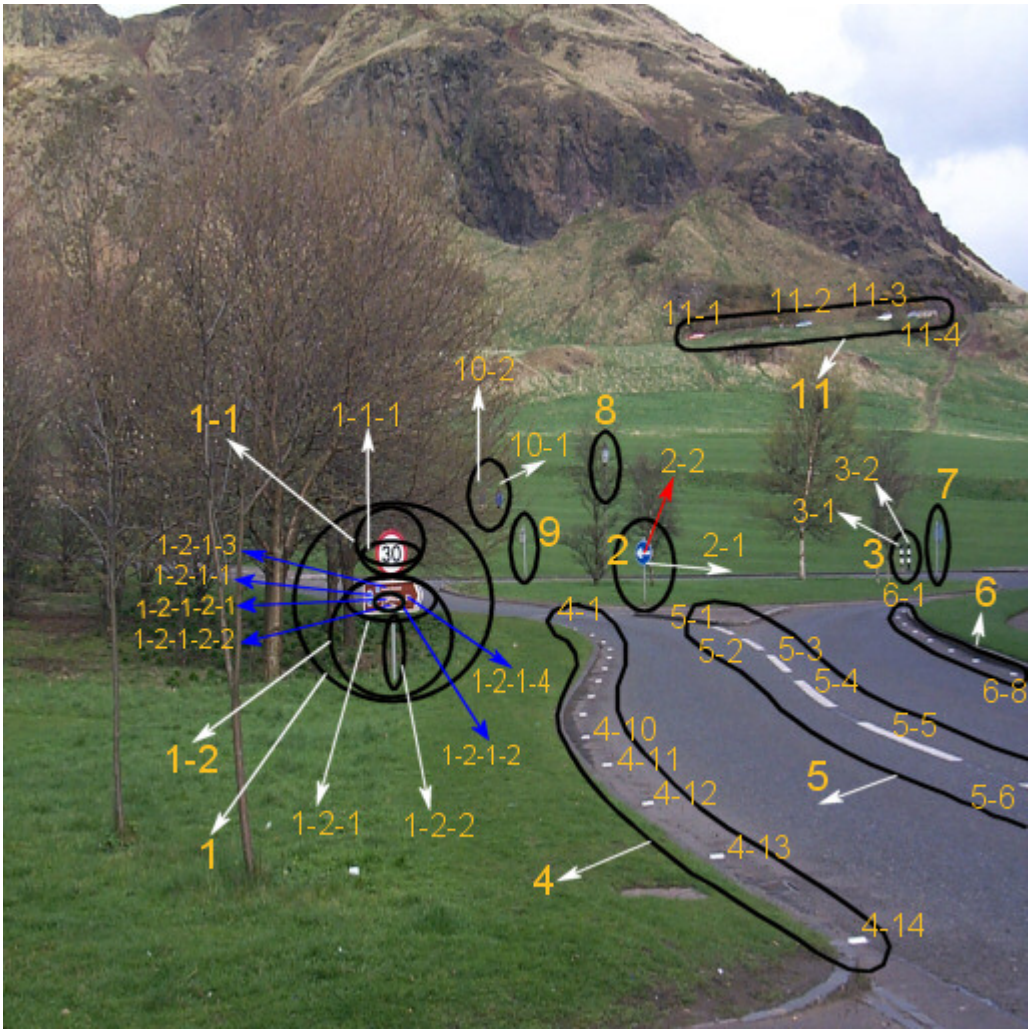

Figure 4.11: The identifiers of groupings in the given photograph.

Figure 4.12: The intensity pyramid built from the photograph given in Figure 4.10.

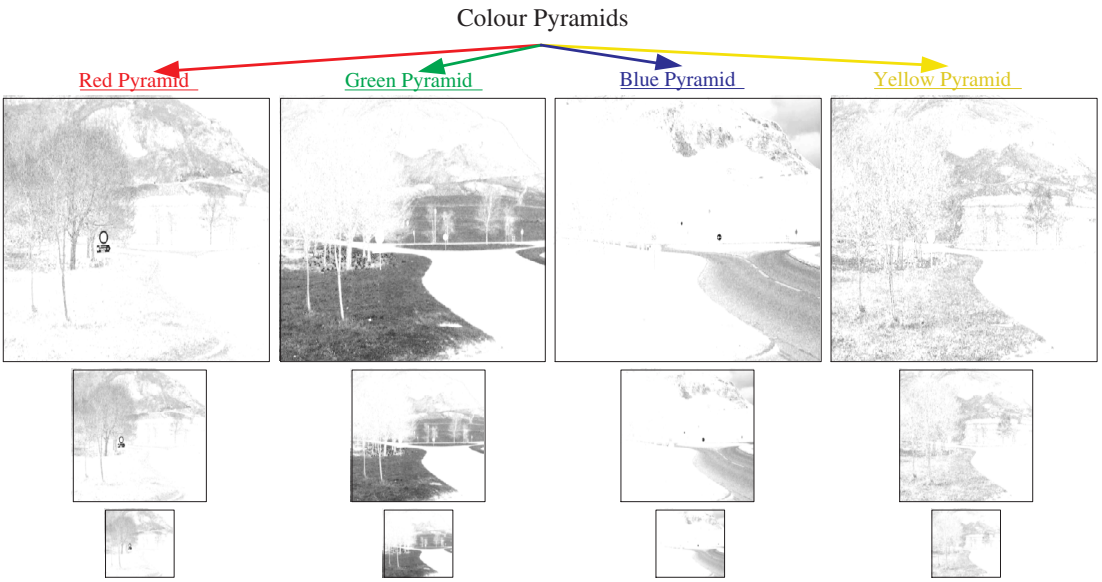

Figure 4.13: The four colour pyramids built from the photograph given in Figure 4.10. (The graphs are black-white inverted to improve visibility.)

because the model theory is based on object attention in which a grouping competes for attention using its entire salience integrating the strength of all its components. Thus, the saliency shown here is the grouping salience rather than that of each pixel within the grouping. However, the grouping-based computation approach can also be applied for spatial attention if each pixel is considered as a grouping. Figure 4.18 gives the saliency maps obtained from the same outdoor scene for individual pixels at the coarsest resolution (graph C), middle resolution (graph B), and finest resolution (graph A). The $1/\rho$ parameter for Gaussian weighting for this experiment is set to 2%.

According to the obtained results, the order of attention shifts is shown in Figure 4.19. It can be seen, the attention movements basically coincide with the salience difference between the objects in the scene. Some groupings, such as grouping 6, which consist of several very small sub-groupings, do not exist at the coarser resolution. They either have no way to take part in the competition, or lose much support from their smaller members or components or from their surroundings which may be useful to compete for attention at the finer resolutions. So generally, they lost some possible advantages when at the finer resolutions.

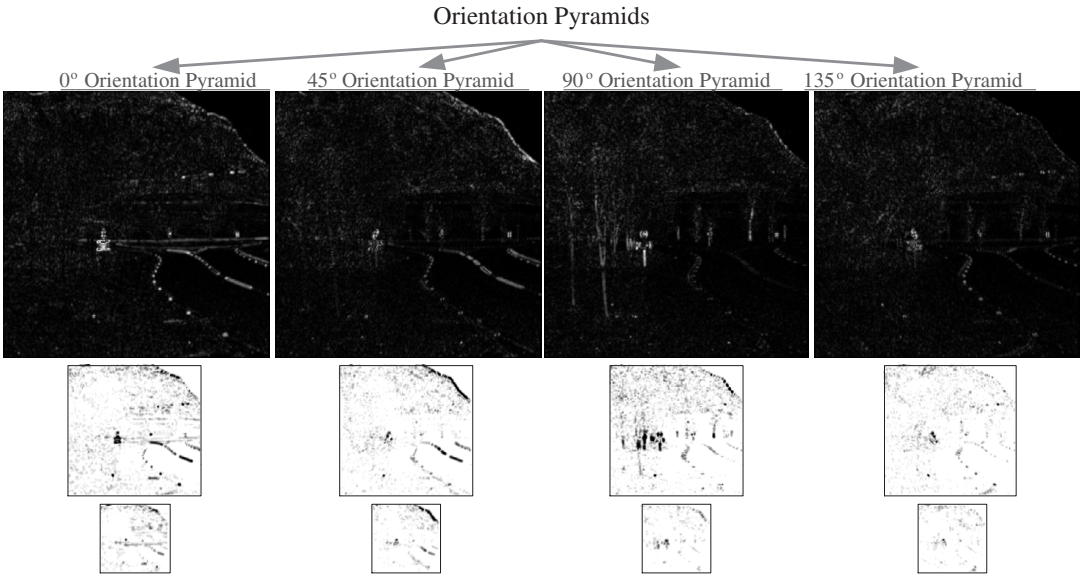

Figure 4.14: The four orientation pyramids built from the photograph given in Figure 4.10. (Graphs in the second and third rows are black-white inverted to improve visibility.)

4.1.2.3 Hierarchical Selectivity From Far to Near

Three colour images shown in Figure 4.20 are taken using different resolutions from far to near distance (64×64, 128×128, and 512×512) for the same outdoor scene. The scene is segmented (by hand) into 6 top groupings (identified by the black colour numbers: one object grouping 6 and five regions here) and 5 of them are hierarchically structured except grouping 4. In the coarsest image, only grouping 6 (one boat including two people) can be seen. In the finer image, sub-groupings 5-1 and 5-3 within top grouping 5 appear but they lose details at this resolution. The smallest boat (i.e. sub-grouping 5-2 of grouping 5) can only be seen at the finest resolution. The saliency maps of groupings during attention competition are also briefly shown in Figure 4.20 where darker grey shades denote lower salience.

The competition first occurs among the top groupings in the image with the coarsest resolution. The most salient grouping 6 therefore gains attention. When giving a "yes" to the top-down attention setting ("view details" flag), attention will shift to the sub-groupings of 6. Two people and the boat then begin to compete for attention. If a "no" is given or after grouping 6 is attended, attention will shift to the next winner grouping 2. If a "yes" is given to the "view details" flag of grouping 2, attention will first select sub-grouping 2-1 and then shift to sub-grouping 2-2. After attending 2-2, if continuing to view the remainder of grouping 2, attention will shift to the finer

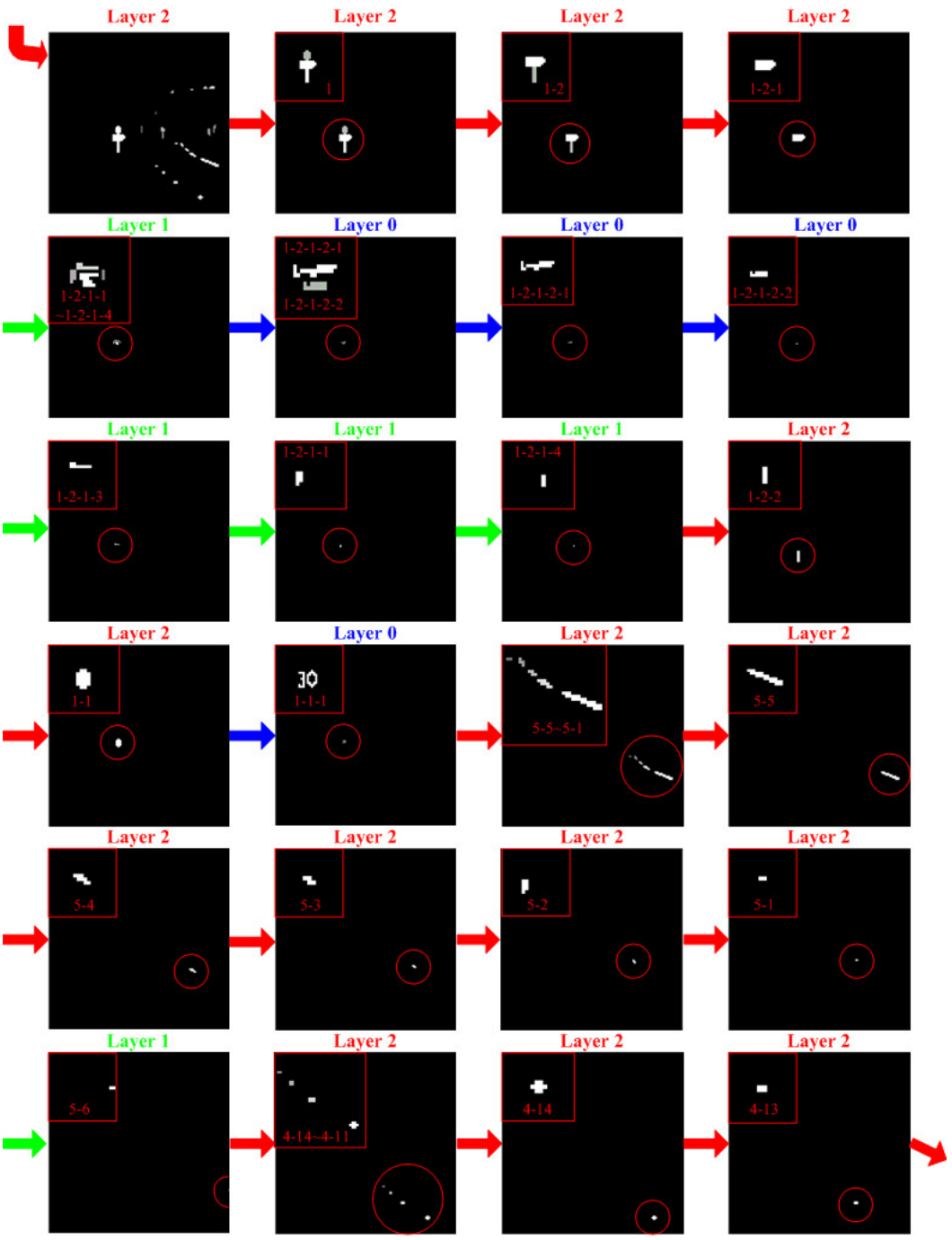

Figure 4.15: Salience of the attending grouping and competing groupings, as well as the sequence of attentional movements. The red, green, and blue arrows denotes that attention is at or switched to the coarsest resolution, middle resolution, and finest resolution respectively. The small red panel at the top left corner in each slide shows a zoomed view of the objects. The red circle/semi-circle indicates the focus of attention. The grouping identifiers are also given in each panel.

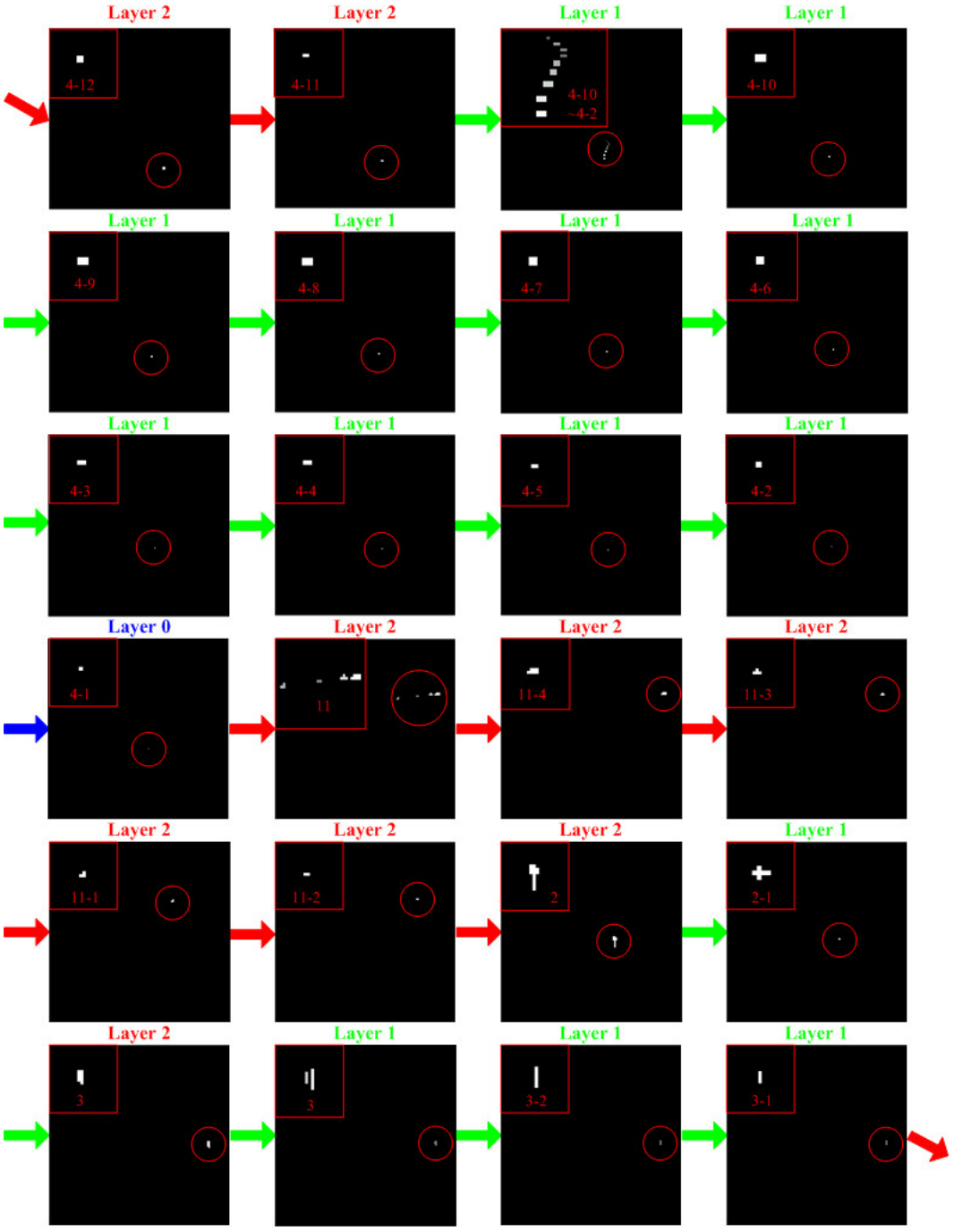

Figure 4.16: Continued slides of Figure 4.15

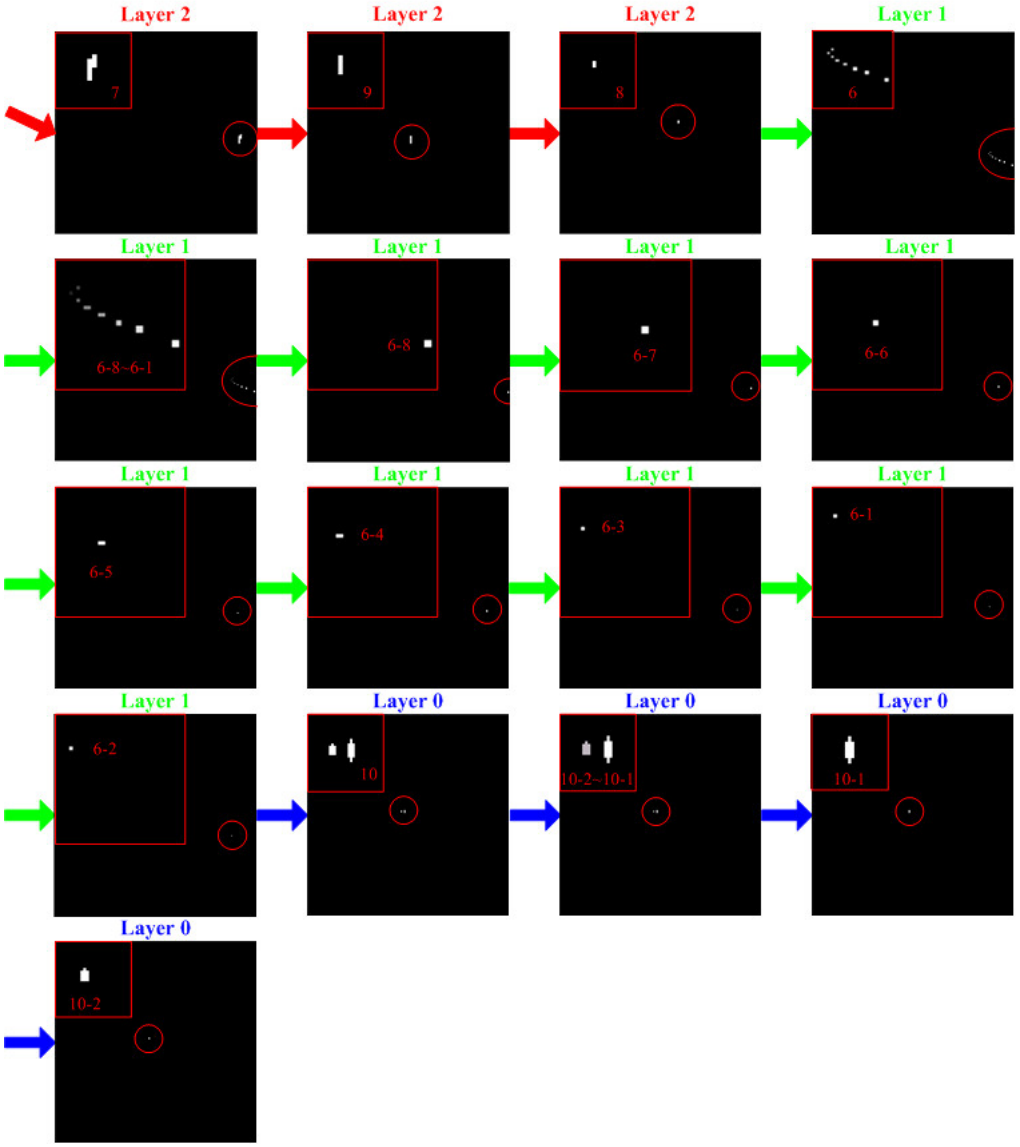

Figure 4.17: Continued slides of Figure 4.16

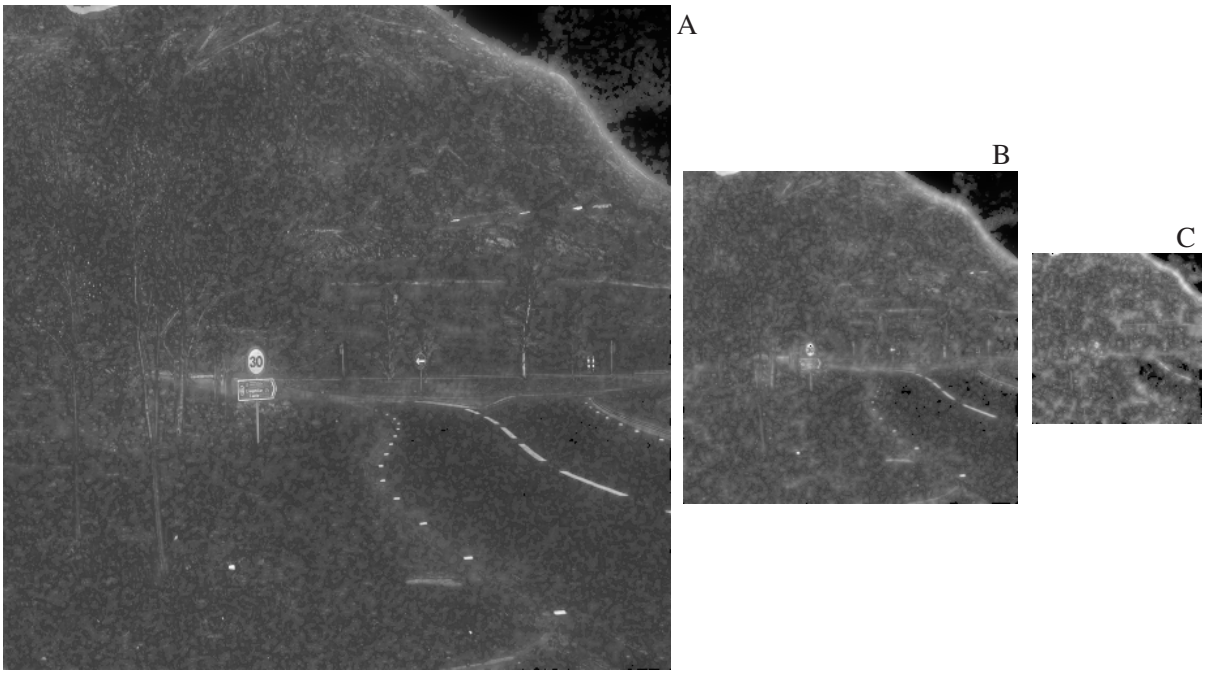

Figure 4.18: Applying the model for space-based attention. Each pixel is an individual grouping. Only the raw saliency maps of pixels at three resolutions are shown here in shades of grey.

resolution to visit 2-3. When grouping 5 is attended, the lake (excluding grouping 6) is visited first and then attention shifts to the finer resolution scene where boats 5-1 and 5-3 start to compete for attention. In the case of giving a "yes" to the top-down flag of the winner 5-3, attention will shift to the finest resolution scene to check its details. Then attention goes back to the previous finer resolution scene and shifts to 5-1. After that, attention shifts again to the finest resolution scene. Thus the smallest boat 5-2 at the finest resolution is attended.

Figure 4.20 shows the overall behaviour of the model performed on the scene. Using this same scene, when stronger and stronger noise was added above $\sigma = 17$ for Gaussian noise, the order of the attention movements changed. The above results clearly show hierarchical attention selectivity and believable performance in a complicated natural scene. In addition, although this model is aimed at computer vision applications, the results are very similar to what we might expect for human observers. The attention movements shown in Figure 4.21 reveal the reasonable shifts of visual attention for this natural scene.

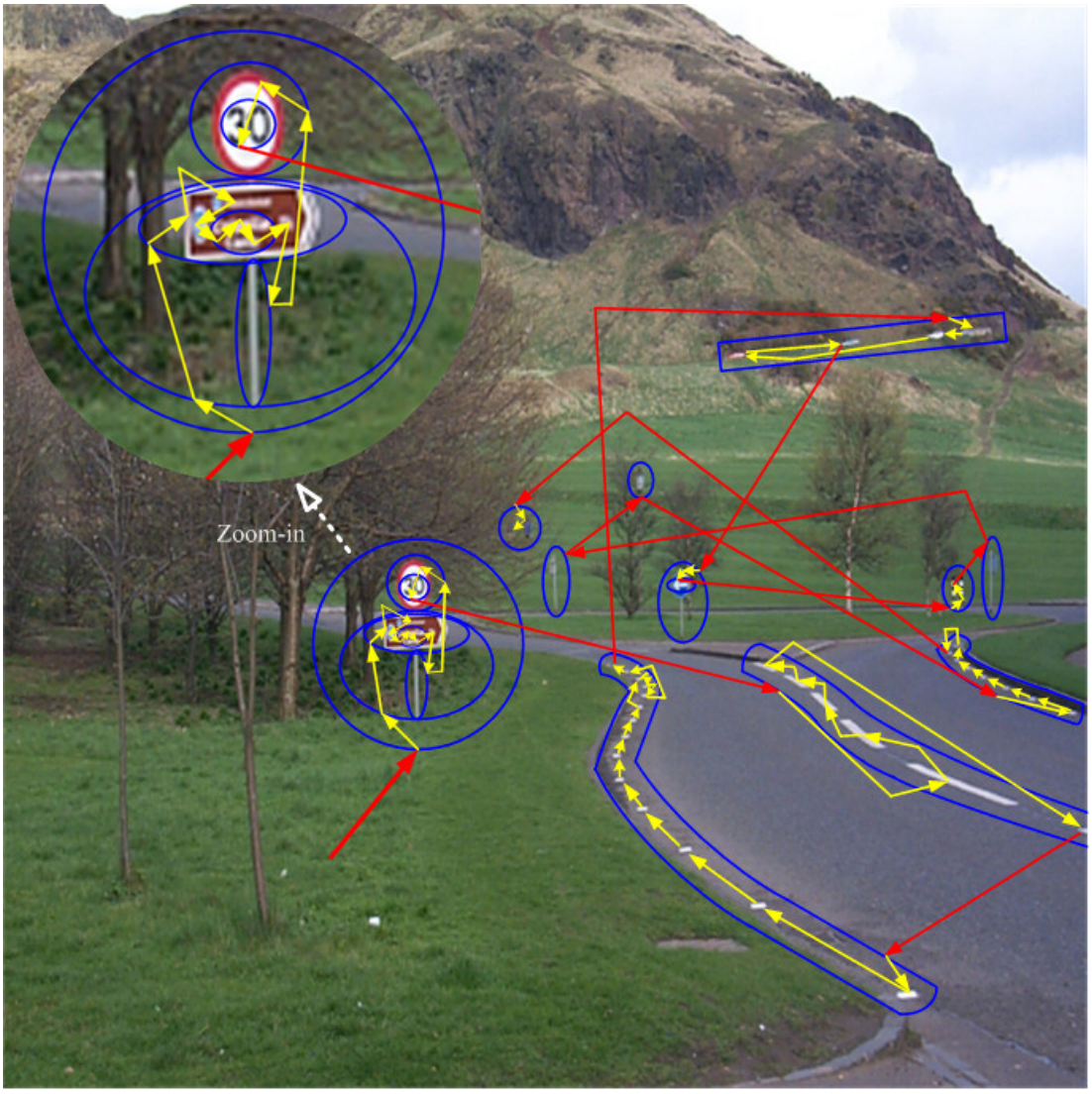

Figure 4.19: The overall trajectory of attentional movements of the model in multiresolution. Red arrows show attentional shifts from one grouping to another. Yellow and purple arrows show attention switches within groupings. The circles denotes the locus of attention.

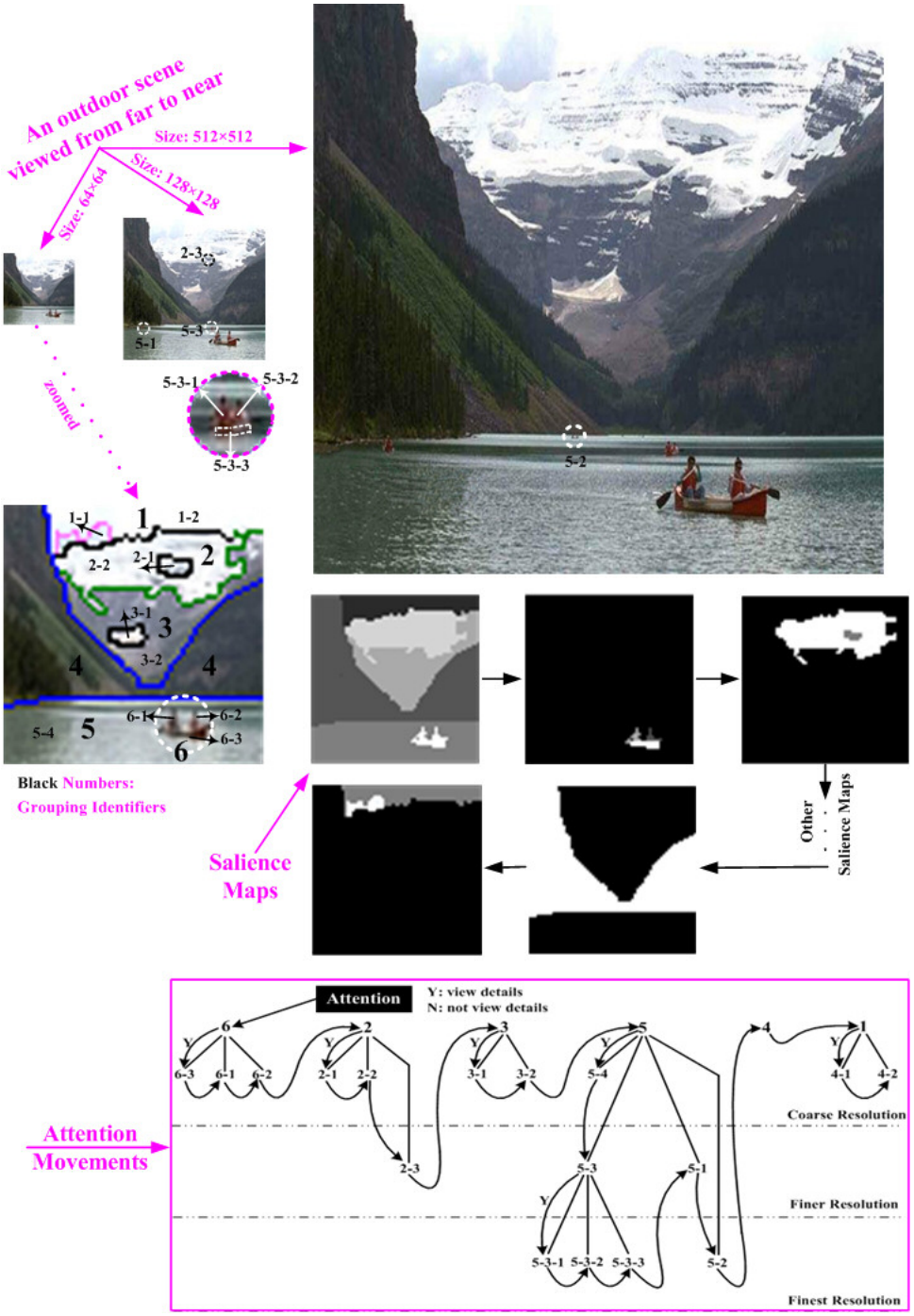

Figure 4.20: An outdoor scene taken from different distances. The dotted circles are used to identify groupings but not their boundaries. The sequence of saliency maps used for each selection of the next attended grouping is shown at the middle. Attention movements driven by hierarchical selectivity is shown at the bottom using a tree-like structure.

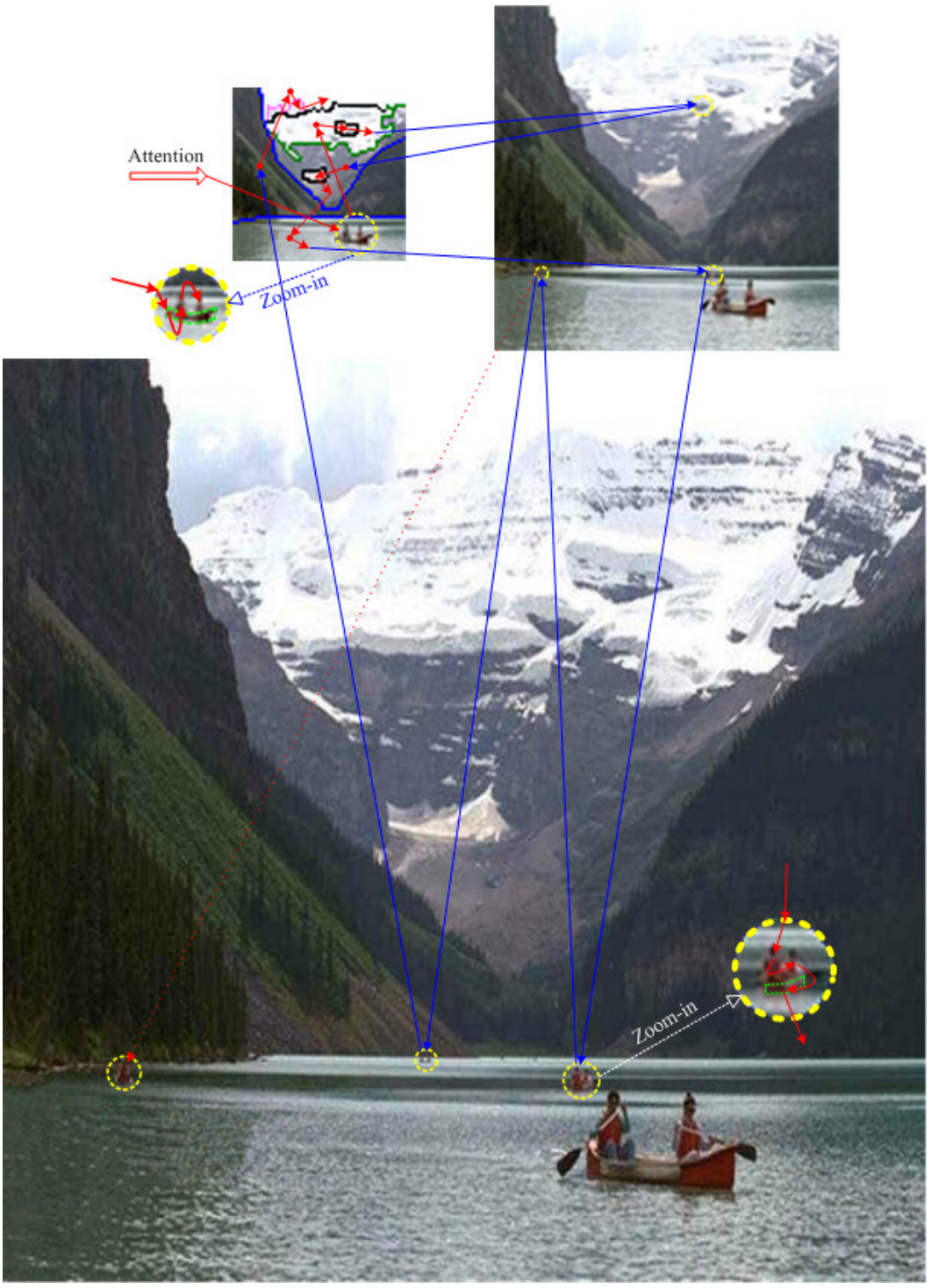

Figure 4.21: The attention movements implemented for the outdoor scene: blue and red arrows indicate attention shifts between and at the same resolutions respectively. Arrows with red solid circles denote attention is attending the top groupings.

4.1.3 Improved Behaviour of Hierarchical Selectivity in Natural Scenes

The previous sections showed the model performance in the complex natural scene. For a complete examination, a positive response was given to each "view details" flag. However, some small stripes (on the road) may be irrelevant to the current visual task and are thus unnecessary to attend in turn. Also some tiny unreadable characters are probably not worth notice by the observer. One possible way to improve the performance on these targets is to incorporate a top-down recognition component or learning process to produce a control function with reasonable salience thresholds according to different environments and visual tasks. The current model does not yet implement this complicated top-down control. Instead, an alternative demonstration of the model's abilities was proposed by using a simple human-computer interaction to give a positive or negative response to the "view details" top-down attentional setting (see Section 3.3.5 for more details).

Figure 4.22 shows a logical diagram of attentional movements working on a hypothetical scene containing three structured groupings. In this diagram, groupings A, B, and C have a decreasing salience order and the left sub-groupings have greater salience than their right siblings. That is, the saliences of A-1, A-1-1-1, B-1, and C-1 are greater than that of A-2, A-1-1-2, B-2, and C-2 respectively. Suppose that attention is currently deployed at grouping A-1-1-1 and a negative answer is given to the check flag of the top-down attentional setting "view details". Then there are multiple (here four) possible destinations of the next attention movement, shifting to A-1-1-2, A-1-2, A-2, or B (as shown in the diagram). In our previous strategy, the most salient sibling of A-1-1-1 (i.e. A-1-1-2) would win the next attention if a positive answer is checked from the "view details" flag of A-1-1. This strategy has advantages of simplicity and following the closest previous top-down setting to the higher level grouping (the parent A-1-1 of A-1-1-1). Here we present an improved strategy for such hierarchical attention shifts.

Suppose $S(X)$ represents the salience of any grouping X. Assume A and B are the most salient of the competitive groupings and $S(A) > S(B)$. Grouping A (or B) has a multi-level hierarchical structure. Then a tree-like data structure can be used to illustrate these structured groupings. Let the salience of the sub-groupings that have the same closest parent be decreasing from the left to the right. Let A-i_1-i_2- $\ldots$-i_j be the current attended sub-grouping at the level j of A (e.g., A-1-1-1 is a sub-grouping at level 3 of A but is a sub-grouping at level 1 of A-1-1 in Figure 4.22). When $i = 0$ or $j = 0$, A-i_1-i_2-$\ldots$-$i_j = A$. Thus the first level sub-groupings of A are $\ldots, A$-i_1, A-$(i_1 + 1)$,

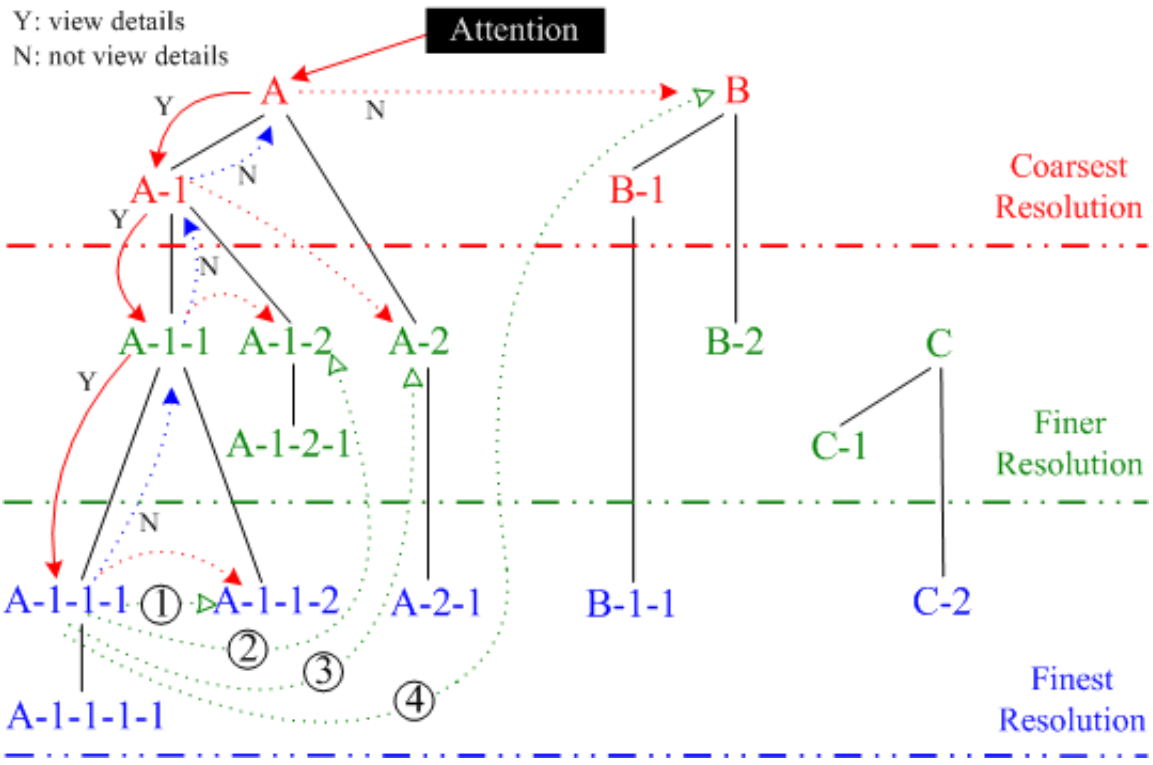

Figure 4.22: Diagram of attentional movements in hierarchical selectivity operating on multi-level structured groupings. Red arrows: attentional movements. Blue arrows: feed-back checker of "view details" flag. Green arrows: possible winners competing for the next attention.

..., the first level sub-groupings of A-i_1 are ..., A-$i_1 - i_2$, A-i_1-(i_2+1), ..., and the rest is deduced by this analogy. Clearly, all sub-groupings left of A-i_1-i_2-...-i_j have already been attended or ignored. A-i_1-i_2-...-(i_j+1) is the most salient unattended sibling of the current attended grouping and A-i_1-i_2-...-$(i_{j-1}+1)$ is the most salient unattended sibling of its parent. When attending A-i_1-i_2-...-i_j, if a negative answer is given to the "view details" flag of top-down attentional setting or this sub-grouping has no child, the next potential winner to gain attention is produced by the following rules:

1. if A-i_1-i_2-...-$(i_j+1) = A$ then attention shifts to grouping B;

2. otherwise attention shifts to the sub-grouping X with salience:
   $S(X) = MAX\{S(A\text{-}(i_1+1)), S(A\text{-}i_1\text{-}(i_2+1)), \ldots, S(A\text{-}i_1\text{-}i_2\text{-}\ldots\text{-}(i_{j-1}+1)),$
   $S(A\text{-}i_1\text{-}i_2\text{-}\ldots\text{-}(i_j+1))\}$

This improved hierarchical selectivity was applied to the natural scene shown in Figure 4.10. Here the entire scene is re-segmented into seven top groupings, as shown

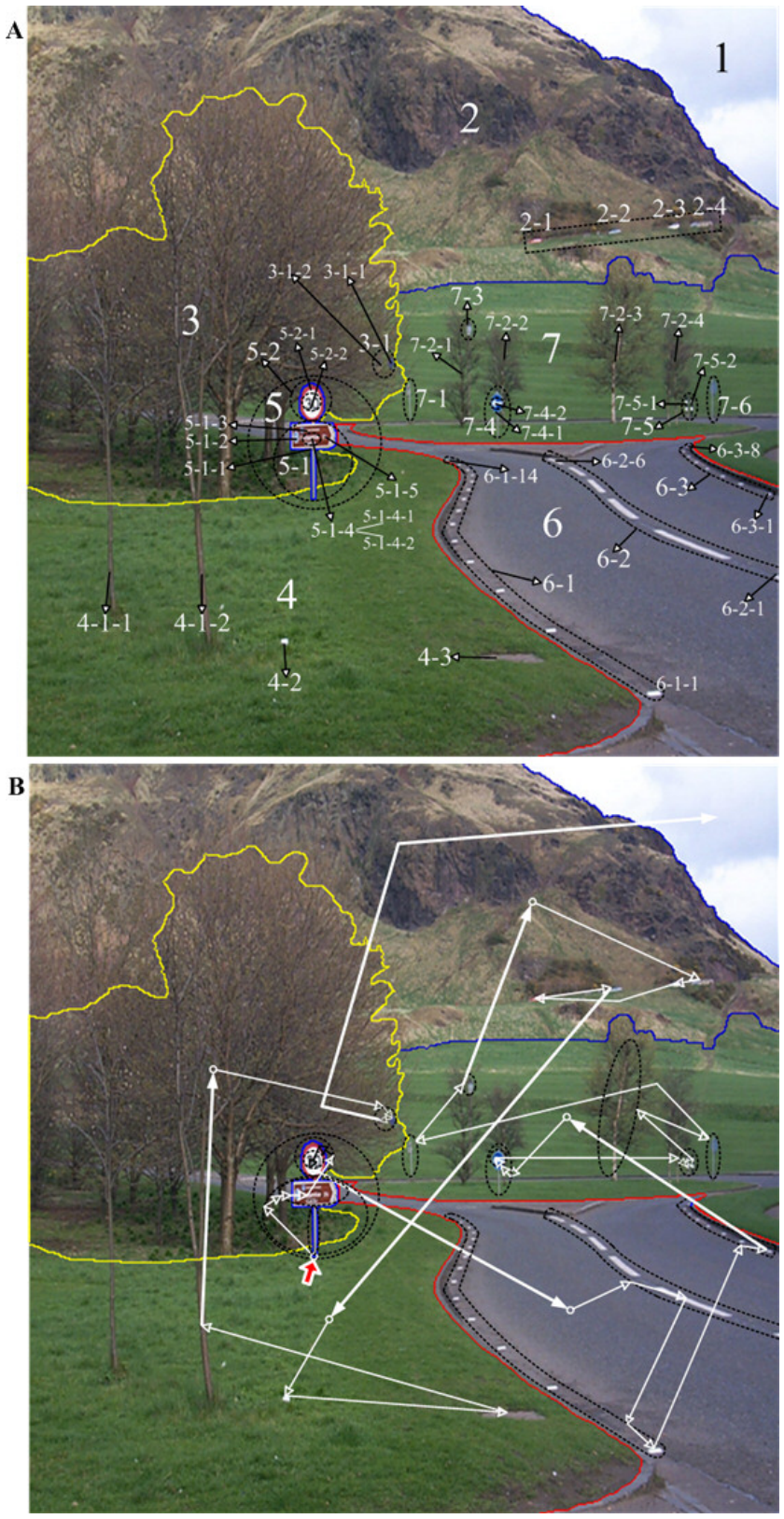

Figure 4.23: A: The groupings segmented from the natural scene; B: overall attentional movements produced by the improved strategy for hierarchical selectivity. Arrows with a hollow circle indicate that attention goes to a top grouping. Air arrows indicate that attention shifts to the sub-groupings. The dotted ellipses are not the sub-grouping boundaries and only used to conveniently show attention movements.

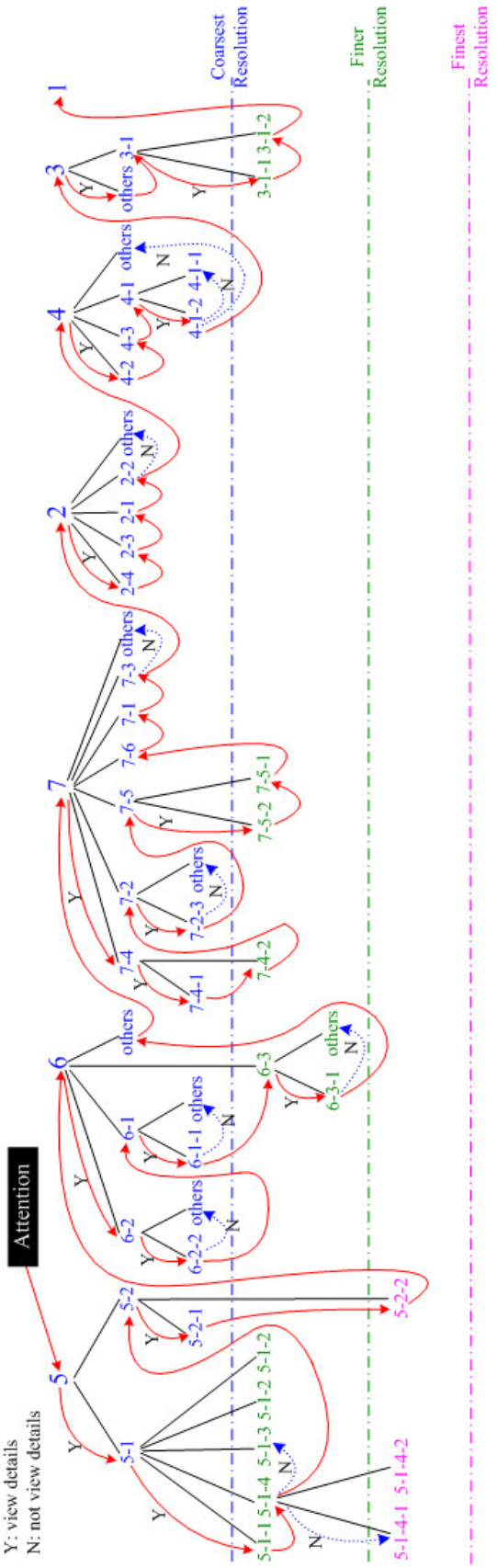

Figure 4.24: The improved performance of attentional hierarchical selectivity on the scene shown in Graph A of Figure 4.23.

in Figure 4.23 (Graph A) by different colour lines. The identifiers of different groupings and their sub-groupings are also given in Graph A. Certain sub-groupings that are segmented within each top grouping are identified and the remainder (such as green grass in grouping 7 or trees in grouping 3) are denoted "others" in Figure 4.24. The "view details" flags of the parent groupings of the small white stripes in the road, trees in the lawns, and some tiny words (and symbols) below the "30" speed limit sign were answered "0" (positive) for the first attending (the first stripe, word or symbol) and "1" (negative) thereafter. Thus most sub-sub-groupings such as those within sub-groupings 6-1, 6-2, and 6-3 of top grouping 6 are also abbreviated as "others" in Graph A, except several first attended sub-sub-groupings (for example, grouping 6-1-1). The $1/\rho$ parameter of Gaussian weighted distance is set to 25% for the global competition between the seven regions and 4% for the local competition within these regions.

Through the improved hierarchical selectivity, more natural attentional movements are clearly seen (Graph B in Figure 4.23). Note here attention is assumed to shift to the center of mass of the attended grouping). The complete hierarchical selectivity procedure for this scene is also shown by using tree-like structure (Figure 4.24) in which the representations have the same meanings as those in Figure 4.22.

4.2 Summary

The mechanisms of object-based and space-based visual attention have been widely investigated in psychophysics and neuroscience research, however, modelling visual attention in computer vision is a quickly growing field, especially for building computable models of covert attention. Until now, to our knowledge, although some computable models for space-based (covert) attention such as Milanese's and Koch and Itti's saliency-based attention models [95, 75, 64] have been successfully built, no computational model for object-based attention has been developed. However, it is worth noting that our work is essentially distinct from the above space-based attention models although saliency map derived from the early work [137, 75] and similar low-level feature extraction are broadly used in many space-based attention models and our work.

Firstly, the saliency map in their work is built upon a location based approach. The saliency map in Milanese' work is actually location-based too and has no difference from Koch and Itti's. In contrast, in our work the saliency mapping is hierarchical grouping based (or proto-object based) and results from the competition between

groupings and their "from local to global" surround rather than locations (or points) and their local surround. Locations are special and primary in the previous space-based attention models but not in our work as already discussed in Sections 4.3.4 and 4.3.6. Secondly, none of the previous research explores a hierarchy of grouping saliency mappings to achieve grouping-based "from coarse to fine" hierarchical selectivity. Thirdly, the most important innovation in our work is that grouping-based competition for visual attention is introduced as a new mechanism so that object-based and space-based attention can be naturally integrated in the same visual selection system. Lastly, thanks to the grouping-based approach, the presented work can naturally make use of and benefit from the effects of various good image (either manual, semi-manual or automatical) segmentation approaches that can not be exploited by the previous attention models. As has been shown, segmentation (also perceptual grouping) is closely linked to visual attention and both affect and benefit from each other [27]. But with those previous attention models (e.g., Itti's and Milanese's models), the segmentation benefit is hard to gain.

This thesis has presented a hierarchical object-based attention model (HOAM) for computer vision. It suggests that object-based and space-based attention can be integrated by using grouping-based competition to deal with dynamic visual selection tasks. By using the integrated competition of proto-objects based on groupings, the selectivity of attention by objects, locations and features can cooperatively work together. In this chapter, the behaviour of the model has been demonstrated on a number of synthetic and real-world natural scenes. The experimental results showed that the model performance concurs with the main findings in the psychophysical literature on object-based or space-based visual attention. Also, the model shows a good performance of selectivity by objects, by features, by spatial regions, and by their groupings on complex natural scenes. Such successful performances depend on three factors:

- *grouping-based competition and saliency evaluation*
- *integrated competition between groupings*
- *hierarchical selectivity*

With the grouping-based saliency mechanism, the pop-out of objects and their groupings can be evaluated in a uniform computational framework. By using hierarchical selectivity to drive attentional movements, the multiple selectivities of objects, features, regions, and their groupings in multiscale resolutions can be performed in

an integrated selection architecture. To our knowledge, the model proposed in Chapter 3 is the first implemented model of object-based visual attention and of integrated object-based visual attention with space-based visual attention in computer vision.

However, there are still several limitations to the current model besides the above strengths. One limitation is that we have not yet built a satisfactory method to deal with the automatic grouping processing. This is a great challenge not only for visual attention but also for computer vision. Another limitation is that we did not present here a complete theory of goal-driven effects on visual attention, which is necessary for understanding visual attention. Lastly, if we use a resolution-varying or retina-like operator at each attention movement, the model will simulate the attention behaviour of human eyes better, because human eyes have decreasing resolution from the fovea to the periphery of the retina. The investigation on the first two points will be done in the future. The work on the last issue will be presented in the following chapters.

Chapter 5

State of the Art of Eye Movements

5.1 Introduction

Complex visual scenes contain a staggering amount of information, more than we can be aware of at one time. If detailed information is needed from many different areas of the visual environment, it can only be obtained by repeatedly moving the eyes so that the relevant objects fall sequentially on the fovea (the highest resolution). Thus, we have to sample visual information by eye movements over time in a series of perceptual actions. Even though a large amount of visual processing occurs in parallel over the expanse of the retinal images, normal visual perception of almost any object or scene is a temporally extended event, as our gaze shifts repeatedly from one object or feature to another. Because those eye movements and fixations take time and result in significantly different retinal images, the visual system must deal with the integration of these even partially overlapping images in spatio-temporal cortex to construct a unified, coherent representation. Eye movements are intimately related to visual (covert) attention. The major functions of eye movements are fixation (to position target objects of interest on the fovea where visual acuity is the highest) and tracking (to keep objects fixed on the fovea during the movements of objects or the observer). Here, the types of eye movements we consider are those which put or keep the targets on the fovea when the head is still.

This chapter reviews the research from psychophysics and machine vision for eye movements, especially on saccadic eye movements. The remainder of this chapter is organized as follows. The next section reviews recent psychophysical research on eye movements for saccade, smooth pursuit, and vergence. This section also discusses the relationship between eye movements and visual (covert) attention. Section 5.3 presents

a number of previous machine vision models of saccadic eye movements. Section 5.4 analyzes the most popular technical approaches for simulating human-like foveal sensing. In addition, the log-polar retina-like sensor employed by HOAF is introduced in more detail. Finally, Section 5.5 summarises some serious principles for modeling human-like saccadic behaviour in machine vision system.

5.2 Eye Movements and Visual Attention

5.2.1 Saccadic Eye Movements

Saccades are rapid, ballistic changes in eye position which take only about 150-200ms to plan and execute. Once a saccade has begun, its trajectory can not be altered. Between saccades, the eyes fixate on the object of interest for a variable length of time so that the visual system can process the optical information available in that location. Most of visual perception occurs during such sequences of fixation actions [12, 83]. Yarbus [152] pointed out that the location and sequence of saccades is not random. Both the physiological and neuropsychological results indicate that the parietal cortex uses information about motor commands to transform visual input from retinal coordinates into an eye-centered representation suitable for the guidance of eye movements [18]. The fact that people do not perceive a moment of blurred vision while the eyes are actually moving (i.e., visual blurring is not perceived during saccades) raises a question – how does the visual system achieve this? The saccadic suppression theory suggested that due to the existence of visual masking, motion during saccades is not perceived because the sharp, clear images from fixation immediately before and after the saccade dominate the blurred images arising during the saccade itself [93] (also see [105, p. 524]). However, there is a lack of evidence from physiology to prove this theory. According to [18], the direct sensory-to-motor coordinate transformation may be an interpretation. The brain must construct a representation to compensate for changes in eye position, and this representation is used to update the remapping of memory trace and spatial attention. The areas of the LIP (Lateral intraparietal area) and the frontal eye fields between which there are strong connections, must work together to construct an eye-centered representation of oculomotor space.

What guides the eye from one fixation to the next, and what determines where and when the eye actually begins to move to the new destination? These turn on the link between saccadic eye movements and visual attention.

5.2.2 Saccadic Eye Movements and Attention

Along with the earlier research on the role of attention in saccades [78, 117, 58], a growing body of studies in modern psychophysics has supported the view that the relationship between (overt) eye movements and visual (covert) attention is a kind of partial interdependence, that is, attention can move freely and is independent of eyes, whereas eye movements require visual attention to precede them to their goal. Correspondingly, the target location and the timing of saccadic eye movements are affected by attention [78, 60, 94]. More recent research on the coupling of attention and saccades has shown that the same spatial attentional mechanism is important for both perception and the programming/execution of saccades. Saccade programming may be directly activated by exogenous cuing, and can also be activated by endogenous processes [60]. Further research indicates that the execution of saccadic eye movements requires focal rather than distributed attention, and this focal attention is guided by a short term memory system which facilitates the rapid refixation of gaze to recently foveated targets [94].

As discussed at the beginning of this chapter and reviewed in chapter 2, human vision usually and mainly uses two kinds of visual mechanisms to deal with complex visual selection tasks in normal visual scenes. In the visual field surrounding the fovea (including the fovea itself), human vision makes use of visual attention – the primary selection mechanism – to scrutinize interesting objects relevant to behaviour without eye movements. When requiring to explore the extended visual environments or potentially interesting objects located at the periphery of the view field, human vision have to employ saccadic eye movements over time to shift the fovea onto these objects for further analysis by attention. This is why a saccade is triggered. With its supporting, visual attention can flexibly accomplish complex selection in a large-scale space. However, this kind of overt shift requires the guidance of attention.

5.2.3 Smooth Pursuit Movements

Pursuit eye movements serve to track a moving object to maintain it stably in the fovea. It is an important function for perceiving the dynamic world. The differences between this mechanism and saccades are [105, p. 524]:

- Smoothness: pursuit movements are usually continuous and smooth unlike jerky or abrupt saccades, although to a certain extent, they can be jerky for tracking a non-smoothly moving object;

- Feedback: pursuit movements need sustained information feedback from the pursued object for modulation in time;
- Speed: the maximum speed of pursuit movements is about 100 degrees per second, slower than saccades. Successful pursuit (i.e., a smooth and exact pursuit) depends on the ability of the visual system itself, as well as the speed and trajectory of the tracked object. Importantly, some studies point out that practice can improve dynamic visual acuity for better pursuit [90].

5.2.4 Pursuit Eye Movements and Attention

Some researchers have explored whether pursuit eye movement requires visual attention for deciding which object is chosen for pursuit [43, 74, 76, 77]. Their research results have shown that the oculomotor subsystem of smooth pursuit indeed receives an input from the voluntary attention system and there exists a single attentional mechanism shared by perception and pursuit eye movements.

5.2.5 Vergence Movements

Vergence movements serve to fixate an object moving in depth and results in a perception of depth by locating the object on the center of the fovea through two eye rotations. Near objects produce strong convergence and far objects cause little or no convergence. Vergence movements are slow, rarely exceeding 10 degrees per second [105, p. 525]. They differ from pursuit mainly in that vergence movements are disconjugate (eyes rotate in different directions at the same time) rather than conjugate (both eyes rotate in the same direction at the same rate) like binocular pursuit movements. If an object motion has both depth and direction, then these two eye movements work together for tracking it accurately.

5.2.6 Vergence Movements and Attention

Like saccadic and pursuit eye movements, vergence movements appear to be under control of voluntary covert attention for target selection, especially displaying attention into different locations in stereospace [59, 40].

5.2.7 Conclusions of the Relationship Between Eye Movements and Attention

In summary, it is important to make a distinction between visual selection accomplished by visual attention and eye movements though eye movements and shifts of attention usually seem to be closely tied [106, p. 80-88]. Visual attention can work independently of eye movements. Moreover, (covert) attention plays a critical role in guiding programming and executing accurate movements including saccades, smooth pursuit and vergence movements which are preceded by shifts of attention. With the help of eye movements, attention can select interesting objects in a complex visual environment more flexibly and efficiently. The relationship of eye movements and selective attention is mandatory regardless of whether eye movements are triggered exogenously or endogenously (i.e. by visual stimuli or by internal factors) [60]. Visual (covert) attention moves more quickly than overt eye movements so as to check a potential fixation point for next eye movement. Kowler et al. [78] found that the amount of attention required for production of an accurate saccade is rather modest but drawing too much attention away impairs saccadic latency, accuracy or both. Interestingly, McPeek et al. [94] furthermore found that the priming of "pop-out" (such as repeated target color) shortens saccadic latency and improves accuracy. Attention may be thought as the primary mechanism of visual selection with eye movements playing an important but supporting role [105, p. 570].

5.3 Conventional Machine Vision Models of Saccadic Eye Movements

There have been many saccadic models developed for machine or active vision, but the majority of them focused on modelling the pure saccadic eye movements and completely ignored visual attention or took them as the same mechanism. The reason behind this may be related to the visual environments, since active vision systems may be designed for specialized visual tasks rather than for biological plausibility. Also, in some visual environments, active vision can possibly employ other selection mechanisms to solve special problems without visual attention, but with extra costs [3].

In the work of Sela and Levine [125], a log-polar representation is used to sample the peripheral images and the interest points are defined as the points of intersection of lines of symmetry between edges in the gray-level images. The authors claimed

their system possesses a near real-time performance. The model proposed by Rybak et al. [120] contains a low-level subsystem to perform fovea-like sampling and primary feature (edges) detection, and a high-level subsystem of "what" (sensory memory) and "where" (motor) structures. The novelty of this model is to achieve invariant representation and recognition by attaching the feature-based reference frame to the basic edges extracted from the retinal images at the fixation point. These are the typical saccading models that modelled overt saccadic eye movements regardless of visual attention. Many other similar models can be found in the literature [11, 19, 145, 135, 129]. Although some of them were claimed to implement attention mechanisms, it is not true because attention is not equal to overt orienting by eye movements. Attention is independent of eye movements and can freely select objects while the fovea does not move. Visual selection is actually accomplished by attention and eye movements do not constitute selection. It is important to distinguish between eye movements and shifts of attention for visual selection [106, p. 38, p. 53].

Some research integrates the mechanisms of visual attention and saccadic eye movements together. In a model combining saccade and pursuit eye movements [92], dynamic temporal information and attentive cues are considered. Depth and motion cues are used for masking pursued objects and attention can then be directed to a new moving object thereafter required by a saccade process.

Horiuchi, Koch and their colleagues [61] have built a visual tracking system based on analog VLSI circuits, which also involves a saliency map, a winner-take-all (WTA) mechanism similar to their previous work [75], and a hierarchical circuit structure. In the system, temporal and spatial derivatives are used to generate the saliency map which provides the input to the WTA and direction of motion. The circuit at the selected location signals the position for the saccadic system to foveate the target, and the motor information for the smooth pursuit system to match the speed of the target.

Recently, a system of data- and model-driven gaze control, which includes salience-based bottom-up and knowledge-based top-down interaction with behaviour component was proposed by Becker et al. in [3]. The low-level processes are implemented on a neural basis and the symbolic processing is employed for the higher stages. In their model, the gaze shift is decided by the behaviour controller and a saccade is always directed toward the location of an activity cluster stored in the object file.

However, the above models did not distinguish attentional selection and overt saccadic eye movements. It is unclear in these models how (covert) attention acts as the primary mechanism to carry out visual selection nor how covert selection and overt

saccading shifts work cooperatively at different levels to perform complex visual selection in a large-scale spatial visual environment. Furthermore, these models only involved space-based attention mechanisms and did not take object-based attention into account at all. Thus they may not provide effective visual selection under general natural scenes in which object-based visual selection is inevitably required [122, 133].

5.4 Space Variant Sensing

5.4.1 Introduction

Human vision makes frequent use of discrete saccadic fixations to explore the visual world and to help visual attention quickly gather interesting information critical to current behaviour. Human eyes use smart foveated imaging to nonuniformly sample the visual world and to intelligently control available multiresolution information across the retina of the eye. Space variant sensing or foveated imaging exploits the capability of the human retina where the fovea with finer spatial resolution is used to observe interesting objects in more detail and the periphery with increasingly coarser resolution is used to rapidly detect potentially interesting objects in the field of view. This is the clear advantage of employing space variant sensors in machine vision systems.

The retina-cortical mapping in the human visual system can be simulated through a log-polar mapping. Spatial variant sensors implemented in log-polar space provide some important properties such as scale and rotation invariance, wide angle and high resolution viewing without mechanical zooming, and biological plausibility [146, 123]. Log-polar mapping has been widely used for foveation, feature detection, and tracking applications [86, 69]. In the work [25], the optical flow estimation problem is solved by space variant image sampling based on log-polar mapping for achieving high data compression and real-time processing. In the recent work [129], log-polar mapping with receptive fields represented by a vector of modified Gabor filters is used for facial landmark localization and person authentication through saccadic movements. In order to produce a more general formulation of log-polar mapping, the complex-log transformation has been developed [118, 45]. More discussion on space variant sensing can be found in [121].

In our lab, Gomes and Fisher [47, 48] developed a log-polar retina-like sensor which is introduced in more detail as below. This sensor is adopted as the retina-like sensor to sample input scenes in the OADS model which is presented in the next

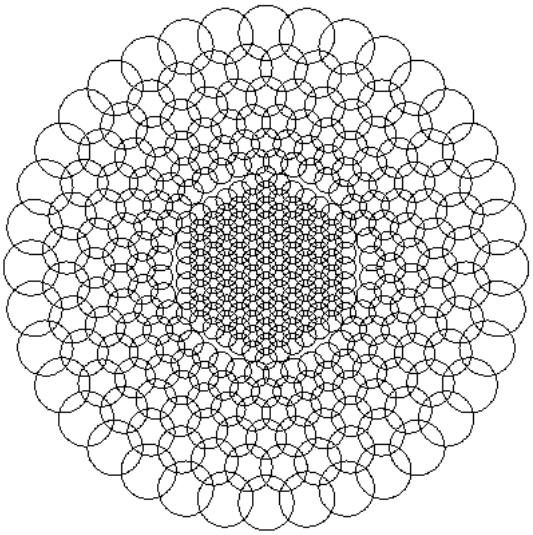

Figure 5.1: Structure of the retinal mask used in the retina-like sensor (reprinted with the author's permission [48]).

chapter for attention-driven saccadic eye movements. More discussion on this sensor can be found in [48].

5.4.2 A Log-Polar Retina-Like Sensor

In the log-polar mapping proposed by Gomes and Fisher, the raw visual input is sampled by means of a retinotopic mask (Figure 5.1) with overlapping circular and normalised Gaussian function receptive fields. The distance from the centre of a receptive field to the centre of the mask is an exponential function of the radius. In the mask, a given receptive field is addressed by two indices: (1) the distance logarithm of the rings to the retina centre and (2) the sector number. The fovea consists of a high density hexagonal receptive field grid. Each retinal layer (or ring) outside the fovea is shifted by half of the angle defining a sector of receptive fields to simulate a hexagonal grid. The radius of the n^{th} outer retinal layer is given by:

$$R(n) = \beta^n R(0) \tag{5.1}$$

where $R(0)$ is the radius of the first layer exterior to the fovea and β defines the geometrical progression of distances of receptive field layers from the retinal centre ($\beta \approx 1.1$). In this way, the radius $r(n)$ of a particular receptive field in layer n is $r(n) = \beta^n r(0)$.

The fovea in the work of Gomes and Fisher was defined as having 11 layers (or rings) of receptive fields with half-pixel radii. Outside the fovea, 37 more layers of

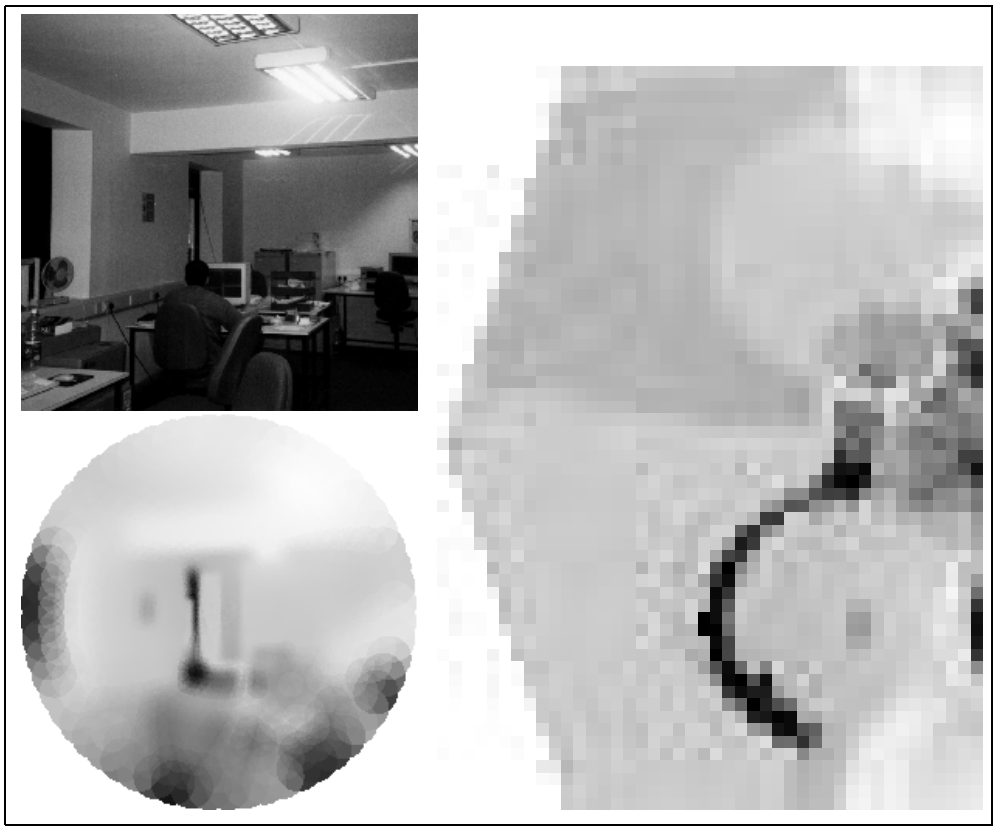

Figure 5.2: An example of deriving a log-polar retinal representation from an input Cartesian image: Top left: original input image; Top right: log-polar image, magnified 4 times; Bottom left: reconstructed retinal image from the log-polar one using receptive fields averaged at intersecting pixels. More details of how to derive a log-polar image from an input Cartesian image and how to reconstruct a foveated (Cartesian) image from the log-polar (retinal) image are given in Section 6.2.

receptive fields are arranged. Each receptive field is approximately overlapped by 60.4% of the diameter. These parameters yield the mask coverage of 256 pixels in diameter. The output O of a receptive field is calculated by:

$$O = log(E) + \sum_{x^2+y^2 \leq r^2} log(L(x,y))F(x,y) \tag{5.2}$$

where $F(x,y)$ is a normalised Gaussian function of the receptive field and applied to points (x,y) in the circular domain of radius r. E is the irradiance falling on an object, and L is the local surface reflectance. They have the following relationship with the intensity $I(x,y)$:

$$I(x,y) = E(x,y)L(x,y) \tag{5.3}$$

The $log(E)$ term gives the receptive field an approximation to the weighted logarithm of the reflectance due to E which is nearly constant over local image regions. When combined with a feature extraction process based on the convolution of a zero-sum mask, this log-polar sensor can help achieve some level of illumination independent feature extraction. Figure 5.2 illustrates an example produced by this sensor.

5.5 Summary

Visual attention and eye movements constitute the complete visual selection process. Based on the review in this chapter, important principles can be drawn for modelling attention and eye movements in the same machine vision system:

1. Attention is essential for eye movements:
   This is the most important rule to model a biologically plausible as well as effective system of eye movements. By following this rule, our approach can solve the problems: What controls the gaze shifts and where is the gaze going to next?

2. Attention and eye movements are two distinct visual mechanisms which work at different visual selection levels:
   This is the way the human eye works but has been rarely investigated in conventional machine vision systems.

3. Attention and eye movements work by cooperation and interaction in a dynamic space-time context:
   Human vision system freely uses (covert) attention to select interesting objects

and employs gaze fixations to survey/keep potential attending objects for complex large-scale visual selection tasks. With nonuniform resolution retina sensing and saccadic eye movements, most locations in the visual field of view may be sampled by multiple overlapping spatial resolutions. The salience of an object will accordingly vary spatially across multiple saccades over time. This causes the competition between objects for attention varying in space-time. Depending on current visual behaviour and the competitive dynamics of attention, interesting objects may be reattended more than once. Thus, how to deal with dynamically varying saliency mapping and how to implement the competition for visual selection in a spatio-temporal context must be concerned.

Inspired by the recent findings of psychophysical research on visual attention and saccade, a model for Object-based Attention-Driven Saccadic eye movements (OADS) is proposed in the next chapter, which is the extended version of our previous work (presented in Chapters 3 and 4) within the integrated object-based attention selection framework HOAF and aims to achieve the above principles.

Chapter 6

Modelling Object-Based Attention-Driven Saccadic Eye Movements

6.1 Introduction

This chapter is concerned with modelling human-like saccadic eye movements for the Hierarchical Object-based Attention Framework (HOAF). A machine vision model of Object-based Attention-Driven Saccadic eye movements (OADS) is proposed that implements a general and effective saccading system, especially for the application in real-world natural scenes. Traditional machine vision saccading models either completely neglect the role of visual attention in guiding overt orienting based eye movements, or treat both mechanisms as the same one. Even though some of those models involved visual attention, they were based only on space-based attention. In contrast, the proposed saccading model OADS differs *per se* from those models in terms of its construction theory, architecture, and processing mechanisms, listed as follows:

- Object-based competition for attention-guided saccadic eye movements:
  OADS is built upon our previous work for Hierarchical Object-based Attention Model (HOAM) and together with it to form the whole object-based selection framework HOAF. It shares the grouping saliency mapping and grouping-based competition for object-based attention with our object-based attention model HOAM. This suggests that the competition for visual attention dominates the competition for an overt saccade. Therefore, the fixations of saccades are guided

by attention. Furthermore, because the competition is object-based, a saccade has little chance to fixate on a nonsense or empty destination.

- Two-level selection system composed of visual attention and overt orienting saccadic eye movements:
  Similar to the human visual selection system, visual attention in our work acts as the primary selection mechanism to launch covert shifts and overt foveal movements.

- Covert and overt selection with a time-varying saliency mapping:
  Through temporary inhibition of return, saliency mapping in the field of view varies with covert and overt orienting movements in space-time. Thus parts of the field of view may be reattended/refixated over time.

The above properties endow the proposed saccading model OADS with the biological-plausibility and human-like visual selection behaviour. The work presented here investigates how object-based attention selection can work together with saccadic eye movements to assist visual attention to achieve more flexible visual selection in a large-scale visual space, and how the two different kinds of shifts of visual attentional selection and saccadic fixation interact and cooperate with each other. Like many other attention and saccading models (e.g., [65, 120, 125]) inspired by space-based attention, the proposed model's theory and performance is also examined in a simulated saccading way, and in static visual scenes by employing a low-level processing of several primary features. However, the current work can in theory be easily incorporated within motor-driven camera systems by integrating obtained multiple saccadic saliency maps (e.g., into a mosaic-like map) but for the application of dynamic scene analysis or smooth pursuit visual tasks it would require motion information.

The remaining of this chapter is organized as follows. The next section accounts for an overview of the saccading model OADS. Section 6.3 introduces the attentional window. Two novel mechanisms of *temporary Inhibition Of Return* (tIOR) and *Attention-Driven Orienting* (ADO) proposed for the model are presented in Section 6.4 and 6.5 respectively. Section 6.6 demonstrates the OADS performance on real-world natural scenes and compares OADS with other work. Finally, Section 6.7 summarises the work presented in this chapter and gives some useful suggestions.

6.2 Overview of OADS

As illustrated in Figure 3.3, OADS mainly consists of four modules working upon HOAM within HOAF: retina-like imaging sensor, attentional window, temporary inhibition of return (short-term memory) and attention-driven directing mechanisms. The space variant imaging uses the log-polar sensor which has already been introduced in Section 5.4.2. This sensor simulates human retinal imaging and has properties: uniform fovea and log-polar periphery, overlapping and Gaussian weighting over receptive fields and hexagonal neighbourhoods. OADS is designed to work as below (see Figure 6.2 for a pictorial illustration):

1. A given scene is first sampled by a random fixation of the log-polar retina-like sensor to create a foveated image;

2. Through the processes from primary feature extraction to grouping saliency mapping on this foveated image, a winner is then generated through the competition pool of attention;

3. A saccade is triggered. The sensor fixates on the winner (its most salient position or centre of mass) guided by the ADO mechanism;

4. A new foveated image is created. A new spatio-temporal saliency mapping is build by integrating all previous saliency mappings over time;

5. Based on the new saliency mapping, visual attention covertly selects interesting objects within an attention window surrounding the fixation point without foveal movement. This kind of covert attentional movements are guided by tIOR until a top level grouping outside the attention window wins the competition by its salience being greater than the total salience of the unattended groupings inside the window;

6. Repeat steps 3-5, a series of foveated images is created by overt gaze shifts. Interesting objects in the whole scene are then effectively selected for processing. During this procedure, some objects may be reattended more than once due to tIOR which yields time-varying saliency mapping for the scene.

7. The scanpath of visual selection is therefore formed by both covert attentional shifts and overt saccadic eye movements.

The processes of low-level feature extraction, grouping saliency computation and competition pool of attention have been implemented in Chapter 3 and 4. The attentional window, tIOR and ADO modules in OADS will be described in the following specific sections. Some important processes involved in the above procedure are discussed in more detail as follows:

Foveated Image

OADS extracts features and further obtains feature maps from the foveated images rather than directly from an original input scene. With the gaze shifts in an input scene, a series of foveated images are produced from the retina-like sensor. This procedure involves a pair of processes: the mapping from the Cartesian input to retinal (log-polar) images and the following mapping from the obtained retinal images to the reconstructed Cartesian (foveated) images [134].

Let n and s be a pair of indices for the ring and sector in the sensor, $V(n,s)$ be a receptive field, the entire field of view can be converted to a log-polar image by just varying the indices n and s within the appropriate ranges:

$$V(n,s) = O(Q(n,s), r(n)) \;\; 0 \leq n < N, \; 0 \leq s < F(n) \tag{6.1}$$

where function O is the receptive field function (as defined in Eq. 5.2), $Q(n,s)$ is the receptive field centre in Cartesian coordinates, $r(n)$ is the radius of any receptive field in ring n, N is the total number rings in the retina, and $F(n)$ is the number of receptive fields in ring n. Given that $R(n) = \text{distance}(\text{Q}(\text{n},\text{s}), (\bar{x},\bar{y}))$ is the radius of a given ring n of receptive fields ($(\bar{x},\bar{y})$ is the centre of the retinal mask), the following constraints must be satisfied in order to make Eq. 6.1 be a log-polar transform in n, s:

$$\begin{aligned} R(n) &= \beta \times R(n-1) = \beta^n \times R(0) \\ r(n) &= \beta \times r(n-1) = \beta^n \times r(0) \end{aligned} \tag{6.2}$$

where $\beta > 1$ is a constant, $R(0)$ is the radius of the first layer exterior to the fovea and $r(0)$ is the radius of receptive fields within that layer.

Because in Eq. 6.1 many input Cartesian intensities under a receptive field are mapped into a single value, this function is not mathematically invertible. But an inverse transform to reconstruct a foveated image in the Cartesian space is more convenient and suitable for an attention mechanism. Thus, by replicating each retinal pixel

all over its reconstructed receptive field area, we have the following transform:

$$W(x,y) = V(n,s) \tag{6.3}$$

where: $0 \leq n < N, 0 \leq s < F(n)$ and $(x - Q_x(n,s))^2 + (y - Q_y(n,s))^2 \leq r(n)^2$. $W(x,y)$ is the reconstructed foveated (Cartesian) image. One problem of this approach is that a pixel in the Cartesian space could be assigned a value more than once because of the receptive field overlapping. A straightforward solution to this problem was to simply average any overlapping pixels in the Cartesian domain.

The parameters of the sensor used in the experiments of this chapter are: 34/66 layers of receptive fields within the fovea, 66 more layers outside the fovea which should be enough to cover the entire Cartesian input image, and a radius of 0.05/0.01 of a pixel for each receptive field within the fovea. A receptive field overlapping of approximately 60.4% of the diameter is defined. We defined the number of layers outside the fovea in such a way that the retinal mask would cover all the rectangular Cartesian image areas and produce not much blurred image periphery so that it is not necessary to introduce a huge bank of different size or size adaptively varying with resolution Gabor filters for orientation detection. The size of the receptive field in the fovea is chosen to be small so that it would be possible to discard the non-log-polar part of the representation.

Figure 6.1 exemplifies the processes described above. For a better illustration, the parameters used to draw this example were different from those used in the experiments in which the foveated images produced are not as blurry in the periphery so that we can directly use the bank of previous steerable filters in our previous work rather than specially designing new complicated orientation filters for log-polar images. The additional reasons to use space-variant sensor and extract features from the Cartesian reconstructed images (rather than directly from log-polar images) are stated as follows.

First of all, using a space variant sensor in this work instead of computing a pyramid of steerable filters (as used in our previous work) at the same fixation point is to take advantage of the fact that the coupling of the attention mechanism with a space variant sensor allows the examination of a hierarchical grouping at different spatial scales without the need of computing all scales at a given foveation because the remaining scales for a grouping will naturally come from the following attention shifts. In the subsequent primary feature extraction, we used the space-variant images derived from a log-polar transform (as described above) together with the bottom level of steerable filters applied to the reconstructed Cartesian images as an alternative (but equivalent)

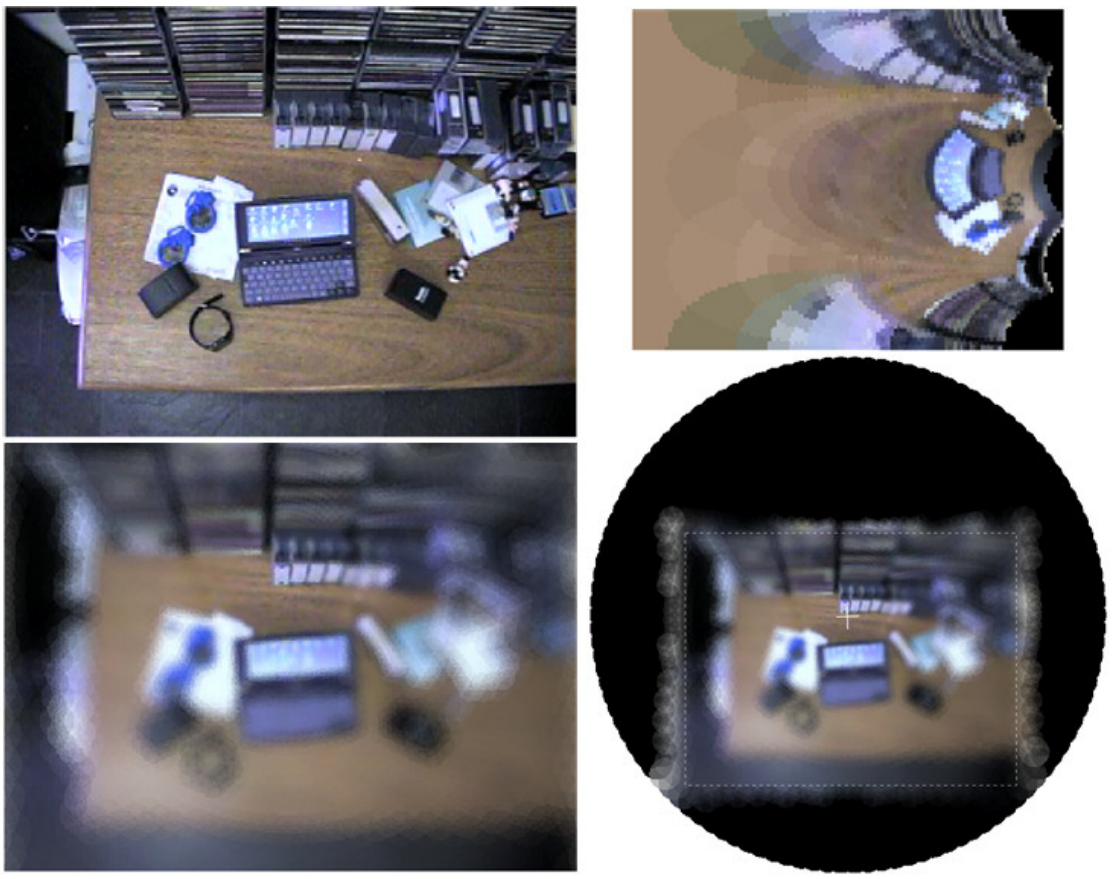

Figure 6.1: An example of how to create a foveated image by the retina-like sensor from an input Cartesian image. Top left: input Cartesian image; Top right: derived log-polar image, magnified for the illustration; Bottom right: diagram that shows the foveation centre (shown by a cross), the retinal mask area (within the large circle) and the clipping rectangle (shown by the dashed white rectangle) that was used to extract the foveated (Cartesian) image (shown in the bottom left) from the usually circular reconstruction, minified for the illustration; Bottom left: reconstructed foveated image in the Cartesian space from the log-polar image (reprinted with the authors' permission [134]).

approach to the space-variant feature extraction. This approach reduces the complexity of the implementation and is faster than the primal-sketch feature extraction for the direct implementation in the log-polar images by Gomes and Fisher [47].

Finally, directly using log-polar images for space-variant feature extraction has some well-known advantages but it also has limitations (e.g., it is necessary to revise or design new image processing operators and it complicates object-based perceptual grouping because the size and shape of image features change radically as the fixation point is moved) [7]. Some possible approaches to overcome the limitations caused by log-polar transforms are to make use of the Mellin-Fourier transform [7], connectivity graph [7], learning-based neural networks [47] and the inverse mapped log-polar image with pyramid algorithms [7] as adopted in this work. One may worry that if the periphery of the reconstructed foveated image is too blurred, it could possibly cause small size Gabor filters difficult to respond properly, so the current saccading model may have limitations. We accept that more general and better primary feature detectors may give better performance in more general applications and will examine this point in the future. But the visual attention process does not care about what is the right or exact value of a grouping's salience. Rather, it mainly considers the relative order of the groupings' saliences. More importantly, the purpose of making use of either a pyramid of steerable filters to construct a pyramidal saliency mapping in our previous work or a space-variant sensor in the current work is to take advantage of hierarchical selectivity of visual attention. This focuses visual resources on the selected objects for more detailed examination and recruits rare resources to quickly and roughly process objects in the periphery by filtering out or eliminating their details, so as to avoid fine processing of the whole visual field. The fine or exact analysis for the peripheral objects is naturally obtained by means of attention shifts. Finally, even if the orientation filters do not work perfectly, this does not much affect the proposed model performance because orientation is only one of the features used for saliency computation and the peripheral objects will all be equally affected so that their relative order in the saliency mapping is not much affected.

The strength of the proposed model results from its grouping-based approach and hierarchical coarse to fine attention selectivity rather than relying on how good the low-level feature extraction is.

Primary Feature Extraction and Feature Maps

Based on an obtained foveated image with a saccadic eye movement (i.e., a shift of the above sensor), OADS uses the same processes as the attention model HOAM's to extract colour, intensity and orientation features and then to build three kinds of (single-layer) feature maps based on the lowest levels (i.e., the highest resolution scales) of the corresponding colour, intensity and orientation pyramids (calculations are given in Section 3.3.3). These feature maps are used to create a single grouping-based saliency mapping for each foveated image. The reason to calculate saliency at only one pyramid level is that OADS adopts space variant sensor to sample an input scene and an obtained foveated image itself contains multiple resolutions which are suitable for the attention mechanism to implement object-based selection from coarse to fine scales. This is different from our previous implementation in the object-based attention model HOAM that employs all pyramid levels and builds a saliency map at each level to achieve visual selection from coarse to fine resolution without using a space variant sensor.

Each foveated image correspondingly produces a single saliency mapping. However, because of the sensor shifts over time, a specific location in the scene will be observed at different resolutions. This means that the saliency of that location may vary in different saliency mappings created from multiple saccadic eye movements and therefore is required to be integrated across multiple resolutions over time. Also, due to the temporal inhibition of return used in attentional selection, the attended groupings have varying saliency over time in a saliency mapping. This requires that the saliency mapping resulted from a foveated image correspondingly adapts over time. A solution for these issues and how a spatio-temporal saliency mapping is built will be described in detail in Section 6.4.1.

Competition Pool of Attention

The competition pool of attention (presented in Section 3.3.5) is used here to generate a winner of the competition for covert attention between the groupings within the attentional window or for a saccade between the groupings (including the parts of the groupings that cross the boundary of the attentional window) outside the attentional window at a certain time. In order to generate the first saccadic shift, all top level groupings in the entire field of view take part in the competition based on the initial saliency mapping obtained from the foveated image by assuming that the initial gaze is fixated at a random position (e.g., the centre) in an input scene. The next saccades

will occur in the areas outside the attentional window when (covert) attention needs to scrutinize a grouping located outside the window. Through this competition pool of attention shared by both covert attention and overt saccading, overt saccadic eye movements are naturally guided by visual (covert) attention.

Transition Between Covert Attentional and Gaze Shifts

Figures 3.3 and 6.2 clearly show that there exist two kinds of shifts in OADS: (covert) attention shifts within the attentional window surrounding the fovea to perform visual selection of interesting objects and attention-guided gaze shifts (i.e., saccadic eye movements) outside the window to help (covert) attention to achieve visual selection in the whole field of view. At any time, it is clear that only one kind of shift may occur. A current shift is made by (covert) attention or by (overt) saccading depends on the current competition and visual behaviour. The nature of the current shift and relevant competition varies over time. If a top level grouping that lies outside the attentional window and wins the current competition when competing with the unattended groupings within the window, a saccadic shift is triggered. Otherwise, when a grouping within the attentional window wins the competition, a (covert) attentional shift occurs. The transition between these two kinds of shifts of (covert) attentional selection and a saccade endows an attention system flexibility to deal with complex visual selection tasks. The implementation of this transition is dealt with by the tIOR and ADO mechanisms explained in Section 6.4 and 6.5 respectively.

6.3 Attention Window

As discussed in Section 2.2, some findings suggest that there exists a relatively sharp boundary between an attended area and its surround. Also, the attended area is movable. The "zoom-lens" metaphor furthermore suggests an attended area of variable size and shape with high clarity at the center and gradually decreased clarity from the center to the periphery [82, p. 27-39]. However, the study on a generally accepted account for the attended area (or called attention window) is still open. Inspired by the above suggestions and for simplicity, a square attention window (with size 256×256 pixels) is adopted in this chapter to show how covert attention shifts and overt saccadic eye movements work together in the proposed two-level object-based selection framework. The attention window is assumed to centre on a fixation point (the fovea) which can be

either the center of mass or the most salient location of an attended/fixated grouping.

6.4 Temporary Inhibition of Return (tIOR)

Inhibition of return (IOR) [109] is a transient bias mechanism which prevents attention from instantly returning to a previously attended location in a short time period. It involves temporal aspects of visual selection. A visual system requires sufficient dwell time to accomplish a visual selection. On the other hand, after a minimum avoidance time, previously selected objects should be allowed to regain visual selection. This is especially useful for a vision system when exploring complex scenes that normally contains hierarchically structured objects that need to be reattended for some further processing. Some findings have shown that there is a close link between IOR and saccade programming [60]. Important evidence also shows that IOR is partly object-based and moves with objects to the new locations [136]. It was reported that IOR can operate simultaneously over several objects, that is, multiple previously selected loci/objects can be inhibited at once [149, 124]. Recent studies have shown that even simple visual tasks elicit many saccades that often repeatedly visit the same objects. And visual attention required by saccades is guided by a short-term memory system that facilitates the rapid refixation of gaze to recently foveated targets [94].

IOR has been broadly used in many computable models of attention/saccade (e.g., [64]) but most of them (including the original HOAM [133]) attend to a target only once without considering the IOR dynamics, i.e., without a reattending/refixation mechanism. However, as discussed above, IOR dynamics is important and necessary for a vision system to deal with complicated visual tasks effectively. The IOR mechanism proposed for OADS is considered in a temporal context and here termed "temporary Inhibition Of Return" (tIOR), which is used to temporarily prevent both attention and saccading from immediately returning to the last accessed object.

6.4.1 Spatio-Temporal Grouping Saliency Mapping

As summarised in Section 5.5, when a scene is sampled by a foveal sensor over time, every location in the scene is observed at multiple resolutions derived from both nonuniform sensing and a series of gaze shifts. Different and partially overlapping foveated images are therefore produced from the fixation movements. Consequently, the saliency of the visual field varies in this spatio-temporal context. The visual system

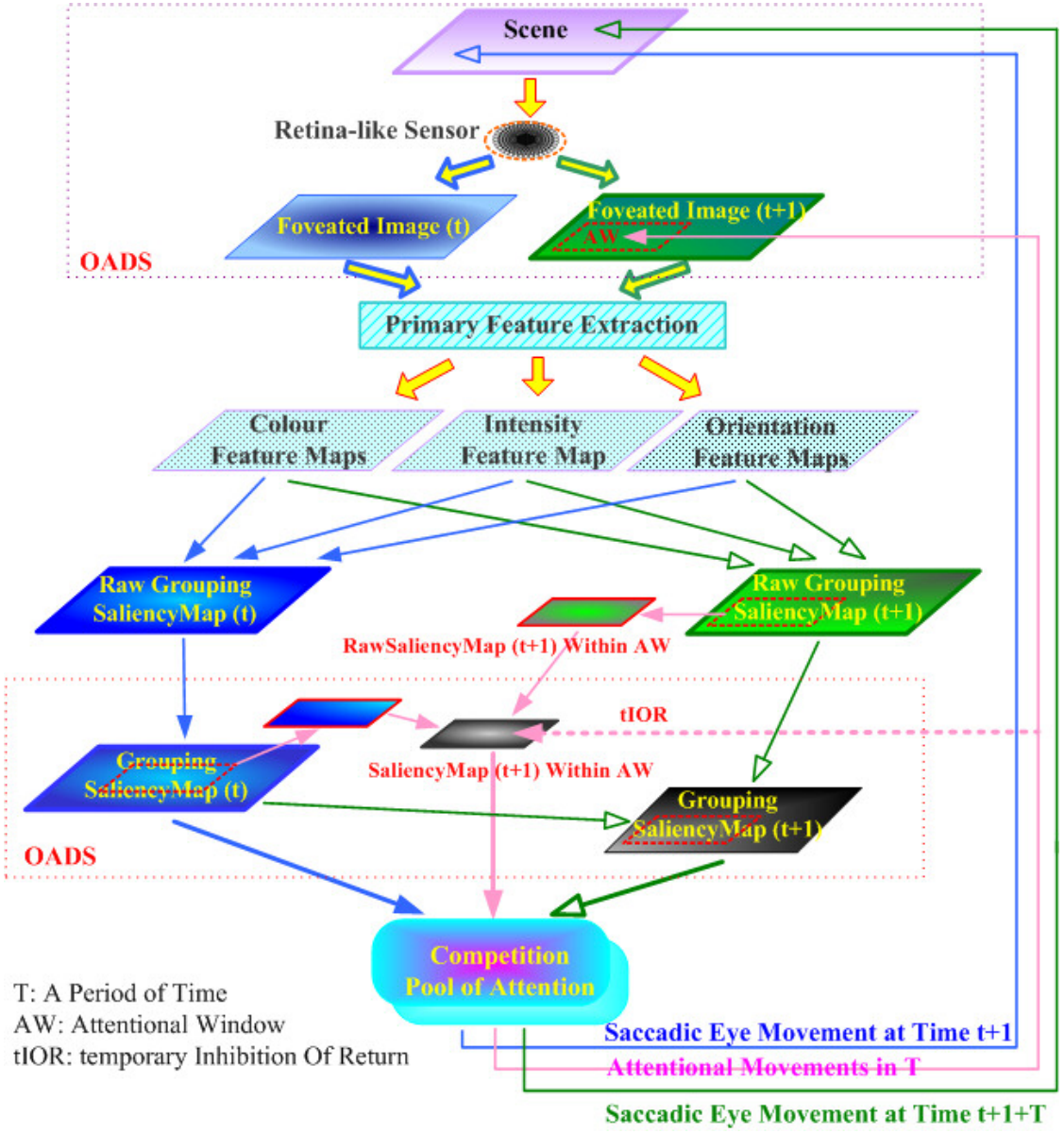

Figure 6.2: The processing flow of how to build the spatio-temporal grouping saliency mapping and how two kinds of shifts of attention and saccading occur within and outside the attentional window in OADS. In the figure, OADS owns the processes within the dotted windows denoted by "OADS" while sharing others with HOAM. Saliency mappings created from the initial fixation at time $t = 0$ (shown by the blue maps) and from the first saccade at time $t + 1$ (shown by the green and black maps) are illustrated here. Green arrows indicate how the next saccade is generated. Red arrows denote saliency calculation within the attentional window (shown by the red dashed frame) which is actually included within the whole saliency mapping. See the context for the further explanation.

must deal with the integration of multiple saliency mappings in a space-time context. The grouping saliency mapping approach presented here provides a solution to this integration.

The approach for spatio-temporal grouping saliency mapping is illustrated by a flowchart shown in Figure 6.2. It is clearly seen that OADS shares lots of processing with HOAM (all modules that are not denoted with "OADS" in the figure). In the flowchart, $RawSaliencyMap(t)$ denotes the (original) grouping-based saliency mapping created by assuming the fovea initially fixating at a random position in the input scene at time $t = 0$. The $RawSaliencyMap(t+1)$ is the (original) grouping-based saliency mapping produced directly from the new foveated image after a saccade shift (i.e., the fovea jumps to a new fixation point) in the scene at time $t+1$. $SaliencyMap(t+1)$ is the new saliency mapping created at time $t+1$ by integrating $SaliencyMap(t)$ obtained before time $t+1$ and the current $RawSaliencyMap(t+1)$. In order to better illustrate how saliency can be combined from the initial foveated image and the next foveated image, time t is assumed as the initial time $t = 0$ in the figure. Thus in this case, $SaliencyMap(t)$ is the same as $RawSaliencyMap(t)$ at time $t = 0$.

Because a saccadic eye movement causes a difference between the two raw saliency mappings at time t and $t+1$ due to multiple resolutions across each location in the scene, the saliency mapping at time $t+1$ needs to be rebuilt to integrate the saliency across multiple resolutions (shown in Eq. 6.4). Based on the recreated saliency mapping at time $t+1$, covert attention shifts within the attentional window surrounding the fovea. When attention needs to select a salient grouping outside the window, the fovea is then directed into that grouping based on the competition between the groupings outside the window. For conveniently illustrating how both attentional and saccadic shifts occur within and outside the attentional window, the saliency mapping within the window is shown in Figure 6.2 separately from the saliency mapping of the whole field of view. But noted that it is actually included within the entire saliency mapping. The only difference is that different Gaussian distance scales are used to compute saliency within and outside the attentional window. A detailed description is given below [1].

The input scene is first sampled by the retina-like sensor to create a (reconstructed) foveated image with space-varying resolution in the Cartesian space. The initial fixation point at time $t = 0$ can be a random position in the scene or at the centre of the scene (as used in the experiments presented in this chapter). Based on this foveated

[1] Here we uses the assumption similar to the one presented in Section 3.3.4, i.e., the input image has already been segmented into hierarchical groupings by any reasonable grouping approach or with human top-down manual help.

image, colour, intensity and orientation features are extracted and then the corresponding feature maps (only the lowest pyramid level is used here for each feature map) are built by using Eq. 3.3, 3.4 and 3.5 (see Section 3.3.3). From the feature maps, an initial raw grouping-based saliency mapping ($rawSaliencyMap(t)$) is created at time $t = 0$ by using Eq. 3.21 and 3.23 implemented in Section 3.3.4.3 and 3.3.4.4. At time $t = 0$, through the "competition pool of attention" (presented in Section 3.3.5), a grouping that wins the competition for a saccadic eye movement between the top level groupings in the whole field of view is generated. The first saccade is then triggered and the foveal sensor shifts to this grouping and fixates on its most salient point or center of mass at time $t+1$. Accordingly, a new foveated image is obtained by the sensor from the new fixation position.

Similar to the creation of $rawSaliencyMap(t)$ at time t, a new raw grouping saliency mapping ($rawSaliencyMap_{\Phi}(t+1)$) is created from the new foveated image at time $t+1$. When creating this new original saliency mapping due to a saccade, the parts of the mapping within (indicated by $\phi = local$) and outside (indicated by $\phi = global$) the attentional window are created separately at the same time. Within the window a small scale Gaussian distance (e.g., $\sigma \leq 5\%$) is used in Eq. 3.13 (discussed in Section 3.3.4.1) which guarantees the competition for attentional selection inclined to a local area. Outside the window a large scale Gaussian distance (e.g., $\sigma \geq 20\%$) is used which guarantees the competition for saccading covering the whole field of view. For the purpose of integrating saliency across multiple varying resolutions and evaluating saliency varying in a spatio-temporal context due to saccadic eye movements, a new saliency mapping ($SaliencyMap_{\phi}(t+1)$) at time $t+1$ is required which is rebuilt upon the current raw saliency mapping integrating all previously obtained saliency mappings before time $t+1$ by using the calculation shown in Eq. 6.4.

Based on this new rebuilt saliency mapping, covert attention shifts over time [2] within the attentional window in the scene to select salient groupings that win the competition for visual attention between the groupings in the window guided by the "competition pool of attention". After each shift of attention, a temporal inhibition (by Eq. 6.5) is used on the previously attended grouping within the window to prevent covert attention from immediately returning to this grouping. Correspondingly, in the saliency mapping $SaliencyMap_{\phi}(t+1)$, the saliency of this suppressed grouping within the attentional window (i.e., $\Phi = local$ here) needs to be adjusted over time.

[2] See Figure 6.2 for an illustration. Attentional shifts within a period of time T are shown by the green arrows.

When a saccade is directed to a salient grouping at time $t+1$ outside the previous attention window at time t, this grouping is also the most salient grouping within the current attention window surrounding the current fovea at time $t+1$. This grouping is consequentially attended and then suppressed. Thus, the same temporal Inhibition Of Return (tIOR) mechanism is actually used for both covert attention and saccadic eye movements. The groupings outside the window and the parts of groupings across the boundary of the window compete for the next saccade guided by the "competition pool of attention".

In Section 3.3.4.4, our previous work on the object-based attention model HOAM uses a pyramidal saliency mapping to implement attentional hierarchical selection of structured objects from coarse to fine resolutions (i.e., from the saliency map built at the highest level of saliency pyramid to the saliency map built at the lowest level of the saliency pyramid) without using space variant sensing in a scene. That is, HOAM uses individual saliency maps at multiple scales rather than a combined saliency mapping across all scales. In the saccading model OADS, as discussed in Section 6.2, because a space variant sensor is employed to sample a scene, each foveated image created from the scene contains multiple resolutions and is suitable for an attention mechanism to perform coarse to fine visual selection. In this case, a specific location of the scene is observed at different resolutions due to multiple saccadic eye movements and accordingly has different saliency when viewed from the different foveal locations over time. Therefore, when building the saliency mapping at the current time, it is required to integrate all of the previous saliency mappings over time. Because at each time a specific location can only cross a single resolution, combining saliency at this location over time is actually equal to combining saliency from multiple resolutions. Based on the above considerations, we have the following approach to build a spatio-temporal saliency mapping at a given time.

Suppose $RawSaliencyMap_{\Phi}(t)$ is the original grouping saliency mapping by adopting the lowest level (i.e., $\hat{\lambda} = \lambda_1$ in Eq. 3.23) of the saliency pyramid created from the foveated image at time t by using Eq. 3.23, $\Phi = global$ and $\Phi = local$ denote the grouping saliency calculation outside and within the attentional window respectively. $SaliencyMap_{\Phi}(t-1)$ is the grouping saliency mapping obtained from the last foveated image at time $t-1$, then the new reconstructed saliency mapping at the current time t

is built as:

$$
\begin{aligned}
&SaliencyMap_{\Phi}(t) = \alpha SaliencyMap_{\Phi}(t-1) + (1-\alpha) RawSaliencyMap_{\Phi}(t) \\
&= \begin{cases} \alpha SaliencyMap_{local}(t-1) + (1-\alpha) RawSaliencyMap_{local}(t) & \text{if} \quad \Re \in AW \\ \alpha SaliencyMap_{global}(t-1) + (1-\alpha) RawSaliencyMap_{global}(t) & \text{if} \quad \Re \notin AW \end{cases}
\end{aligned} \tag{6.4}
$$

where α is a constant $\in (0,1)$, AW indicates the attentional window, $\Re$ is any grouping, $SaliencyMap_{\Phi}(0) = RawSaliencyMap_{\Phi}(0)$. To any time t, $SaliencyMap_{local}(t)$ and $SaliencyMap_{global}(t)$ are contained in the same $SaliencyMap_{\Phi}(t)$. The initially created saliency mapping $SaliencyMap_{global}(0) = RawSaliencyMap_{global}(0)$ at time $t = 0$ is used to generate the first saccade shift from the initial fixation point to a new place at time $t = 1$. Attention shifts occur since time $t = 1$.

The above saliency calculation can be further illustrated in Figure 6.3 which more clearly shows the idea. The calculation shown in Eq. 6.4 provides the temporal integration of the raw saliency across multiple fixations and attentional shifts over time. This can work for both covert attention and overt saccades. As most locations of a scene will be scanned at different spatial resolutions after each saccade, this averaging mechanism can do the integration in a simple way, while responding to the current spatial resolution. After this temporal integration, the competition in the scene across multiple resolutions can be reasonably reflected in the new grouping saliency mapping.

After attention selects an object in the attention window and is going to shift, the attended object must be transiently inhibited so as to avoid regaining attention immediately. As the attentional window moves with a saccade, this inhibition is correspondingly applied to each attended grouping in the whole field of view. Because a fixated grouping is located within the window and is attended first, it is also suppressed when a saccade shifts to it. Thus, the inhibition of return is actually applied to both attentional and saccading shifts.

Let $X = \{x_i, i = 1 \cdots \iota\}$ be a grouping including ι pixels which is currently being attended within the attention window at time t, x_j be an arbitrary pixel in the input scene. Let S_{x_j} be the salience of x_j and $S_{x_j} \in SaliencyMap_{local}(t)$, the suppression function $SUPP(SaliencyMap_{\Phi}(t), X)$ is:

$$
\begin{aligned}
&SUPP(SaliencyMap_{\Phi}(t), X) = \\
&\begin{cases} \underset{\substack{\text{for } i=1 \text{ to } \iota \\ \text{for } j=1 \text{ to } \iota}}{S_{x_j}} = S_{x_j} \cdot (1 - \exp(-\beta(x_j - x_i)^2)) & \text{to all} \quad x_i, x_j \in X; \\ SaliencyMap_{\Phi}(t) & \text{to all} \quad x_j \notin X; \end{cases}
\end{aligned} \tag{6.5}
$$

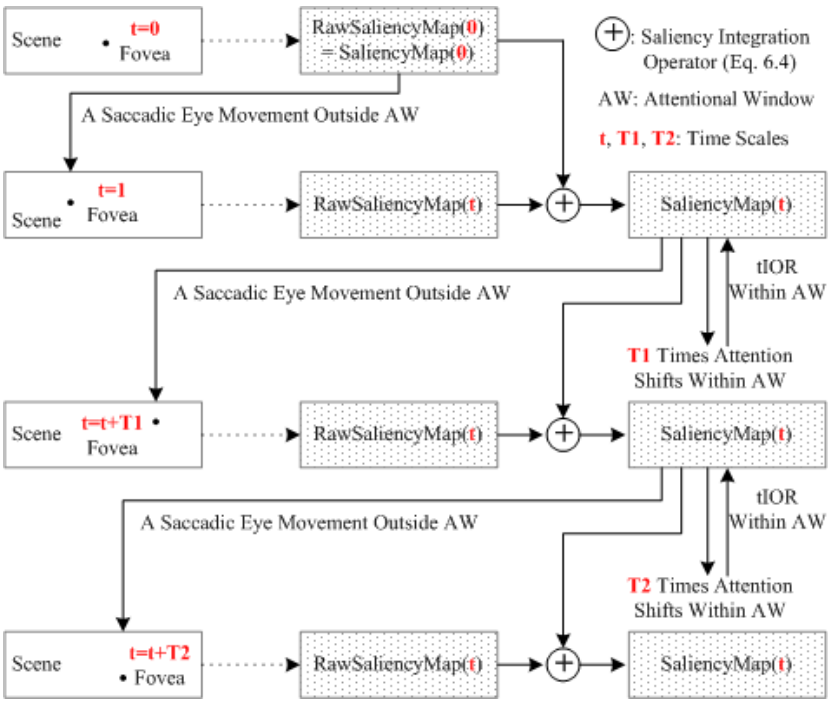

Figure 6.3: Illustration of the calculation of grouping saliency mapping in OADS.

where $\beta > 0$ is a real constant and is here simply set to be $1/D^2$, D is the diameter of X. In the above function, suppressing a grouping X is achieved by taking its each component (x_i) in turn as a suppression center respectively to repeatedly suppress all pixels (x_j here including x_i) surrounding x_i within X. All other groupings are kept no change in the saliency mapping. The above suppressing function can also be made by a simple way that uniformly decreases the entire saliency activities of an attended grouping below a given threshold at one time but we adopt here another way. It is clear that (covert) attention shifts by transiently suppressing the previously attended grouping (i.e., suppressing its saliency activity) within the attention window and a new saccading is triggered to jump out and shift to its new fixation position when the last attended grouping in the window is suppressed.

6.4.2 The Algorithm of tIOR

Given the above spatio-temporal grouping-based saliency mapping within an attentional window, visual attention covertly selects the salient groupings within the window over time. Then, when is the next saccade triggered? That is, when will attention jump outside the current attentional window and shifts to a new salient top level grouping

$t = 0$;

while (the given goal is not reached)
{
create $RawSaliencyMap_{\Phi}(t)$;

recreate $SaliencyMap_{\Phi}(t)$ using Eq. 6.4;

$S_X = Max(SaliencyMap_{global}(t))$ and X is the most salient top level grouping generated by the "competition pool of attention";

saccade to X and create an attention window surrounding the fixation point;

$t = t + 1$; $t' = t$; $i = 1$; $end = FALSE$;

while($i <= n$ and not end) // n is the number of the total groupings within // the attentional window

{ attention covertly selects a grouping i guided by the "competition pool of attention" based on $SaliencyMap_{local}(t')$;

suppress grouping i within $SaliencyMap_{local}(t')$ by Eq. 6.5;

$t' = t' + 1$;

if $S_{in}(n-i) < S_Y = Max(SaliencyMap_{global}(t))$ with $Y \in$ groupings outside the window
{ saccade to Y; $SaliencyMap_{\Phi}(t) = SaliencyMap_{\Phi}(t')$; $end = TRUE$}

else $i = i + 1$;
}
$SaliencyMap_{\Phi}(t) = SaliencyMap_{\Phi}(t')$;
}

Table 6.1: The algorithmic description of *temporary Inhibition Of Return*

located at the periphery of the whole field of view? An ideal solution to this problem will involve complicated top-down guidance and visual and nonvisual reasoning processing. We use here a simple way to achieve this switch from covert attention to a saccade. We assume that within an attention window there are n hierarchical groupings that compete for the covert attention. After a grouping or sub-grouping is attended, it is suppressed by using Eq. 6.5. After attention shifts to the ith grouping/sub-grouping, if a top level grouping outside the attentional window is more salient than the sum of saliency of all unattended $(n-i)$ groupings within the window, attention discontinues the covert selection process and a saccade is ready to launch. We adopt this approach because it can effectively avoid the exhausting search process of attention within an attended area by a reasonably simple way.

Suppose $S_{in}(n-i) \subset SaliencyMap_{local}(t)$ is the sum of saliency of all $n-i$ unattended groupings within an attention window which includes n groupings in total at time t. The other $i \geq 0$ groupings have been attended and suppressed. Let Y be a top level grouping outside the attention window. The algorithm implemented for temporary inhibition of return (tIOR) is shown in Table 6.1.

6.5 Attention-Driven Orienting (ADO)

Saccade and attention are two different principal selection mechanisms in human vision while attention takes the primary role, although they are often intertwined in normal visual tasks. However, most of the previous machine vision saccade systems did not distinguish between these two mechanisms but treat them ambiguously or take them as the same one. Therefore, in these systems, it is not clear how attention works to guide saccading and how these two different mechanisms work together to manage visual selective tasks. This problem has been corrected in OADS.

The attention-driven orienting mechanism built here is specially designed to work in real visual scenes. The competition for a saccade occurs globally among all top level groupings in a given scene, that is, groupings in the periphery of the field of view compete for the next saccading guided by the ADO mechanism. The first saccade is directed to the winning top level grouping which is the most salient one in the first foveated imaging. The fovea moves to the winning position and a new foveated image is created. Correspondingly, the saliency mapping of the scene is rebuilt from the new foveated image. Based on this spatio-temporal saliency mapping, all groupings within the attention window start to compete for covert attentional selection where the fovea

1. Assume the fovea initially fixates on a random position (e.g., the center) of the input scene to create a foveated image by the log-polar retina-like sensor;
2. Create the raw grouping saliency mapping for this initial foveated image. Through the competition pool of attention, a winner among all top level groupings in the whole field of view is generated based on the above saliency mapping;
3. A saccade is triggered and the fovea (sensor) jumps into the winner;
4. Produce a new foveated image at the new fixation point with the retina-like sensor;
5. Create the reconstructed saliency mapping using Eq. 6.4 for this new foveated image;
6. The competition for object-based (covert) attentional selection is triggered within the current attention window;
7. Attention selects groupings over time and suppresses previously attended groupings by tIOR using Eq. 6.5 until a top level grouping outside the window wins the competition by its salience being greater than the total salience of all unattended groupings within the window;
8. Adjust the current saliency mapping over time after applying suppression for each attentional shift;
9. Saccade is ready to shift to a new position;
10. Saccade to the new fovea position (e.g., the most salient point) of the competitive winner created by the competition pool of attention;
11. If (all groupings in the input scene are visited or the given goal is reached)
    Go to step 13;
12. Else go to step 4;
13. Stop.

Table 6.2: The algorithm of ADO

remains fixated at its new foveated location. After processing the covert attentional shifts in the attentional window, the competitive winner for the next saccade is selected from:

1. a previously attended grouping that is the most salient of the groupings that cross the boundary of the attention window and is at least as salient as the most salient top level grouping outside the window; or
2. the most salient top level grouping outside attention window if the above is not available.

Through ADO, a saccade shifts to a new fixation position guided by visual attention. With the help of tIOR, the focus of attention can flexibly jump to check the details of an interesting object located at the periphery to accomplish large-scale visual selection tasks. The detailed mechanism of ADO is described in Table 6.2.

It is clear that saccading is guided by covert attention through the grouping-based competition and grouping saliency mapping. When the fovea is fixated at a position, covert attention shifts freely without saccading in the attention window. When covert attention needs to attend to a new object outside the attention window or the unattended remainder of an attended object that lies across the attention window, ADO drives the saccade to jump out of the current attention window and brings the fovea to the new fixation position. Therefore, the two selection mechanisms of covert attention and overt saccadic eye movements are modelled at two levels while they work together to accomplish complex visual selectivity.

6.6 Behaviour and Performance in Real-World Scenes

A number of natural scenes are used to examine the performance of OADS and to compare the results with those of other research. These applications are described below in detail.

6.6.1 Implementation in Natural Scenes

In scene S1 (512×512) shown at the top left panel in Figure 6.4, there are four top level groupings which are hierarchically structured and consist of several sub-groupings except the rock which is a single grouping. For example, the top level boat grouping

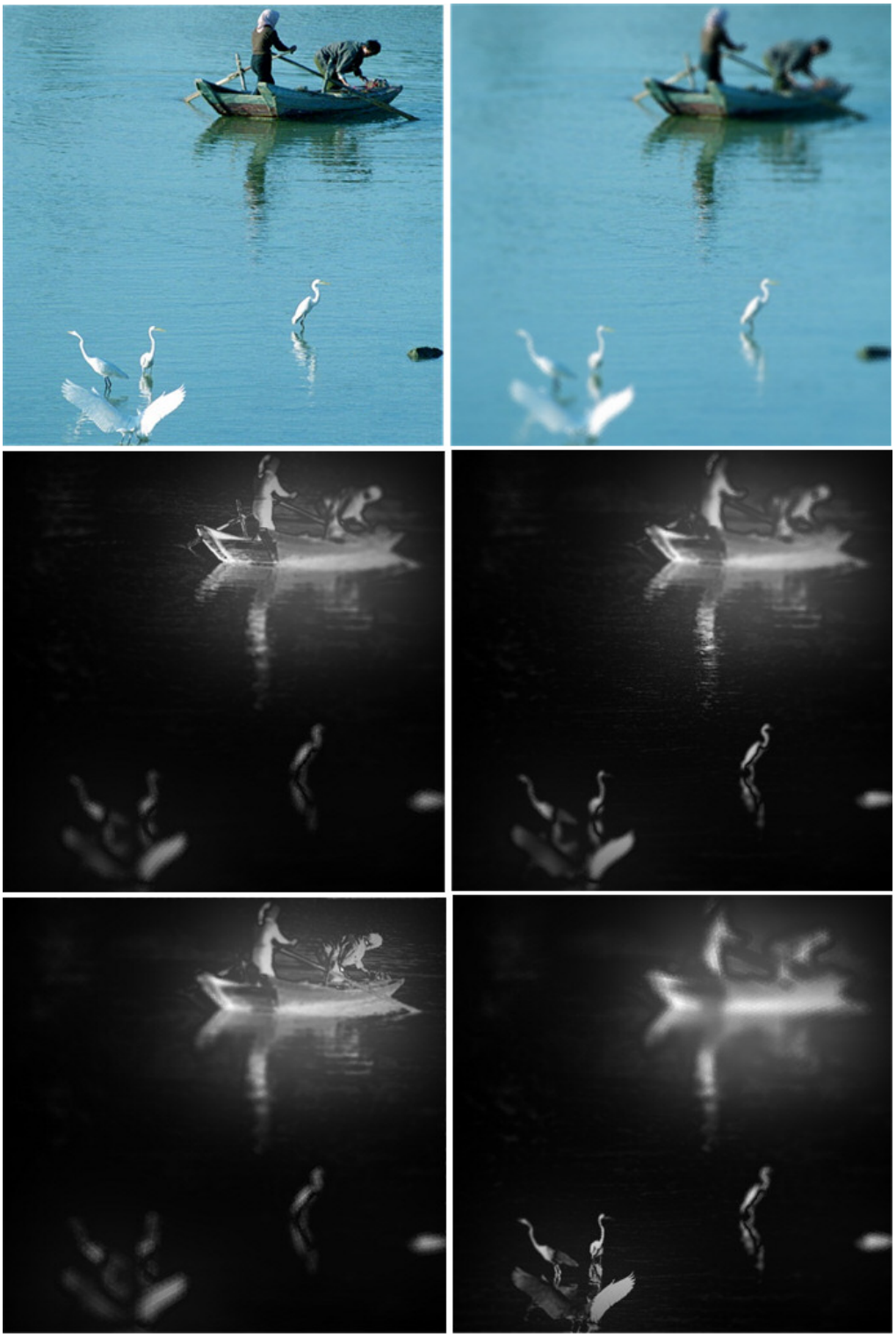

Figure 6.4: Top left: a natural scene S1 taken by a digital camera; Top right: initial retinal imaging; Middle and bottom lines: pixel-based saliency maps computed from the first saccadic eye movement to the one before the last saccadic eye movement.

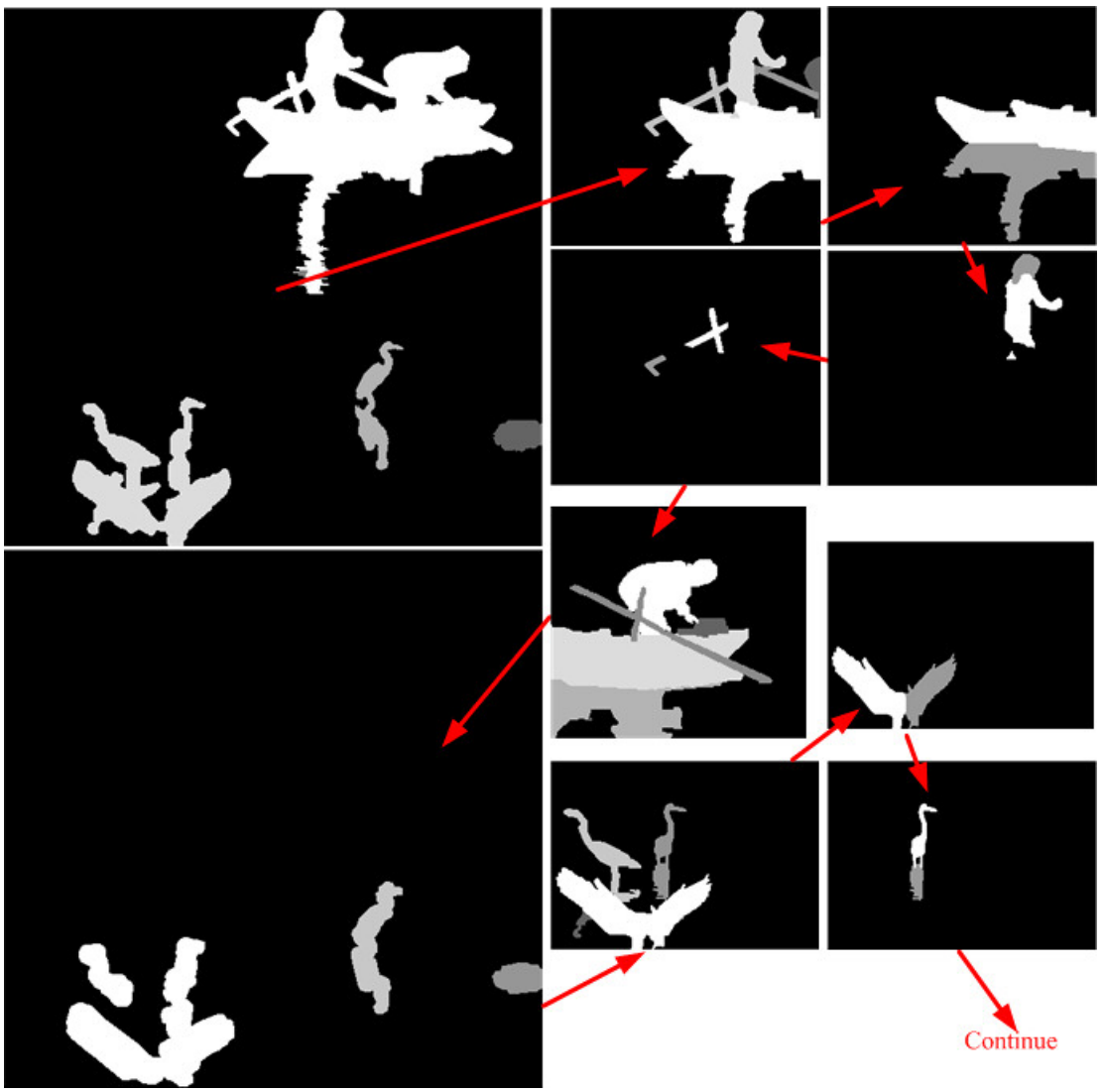

Figure 6.5: Saliency maps obtained from the natural scene S1 during saccadic eye movements and attentional shifts. Different salience strengths of groupings are shown in different grey scales where the brighter is more salient. Large images show saliency maps due to saccades whereas small images show saliency maps within the attentional windows (different window sizes are caused by fovea positions).

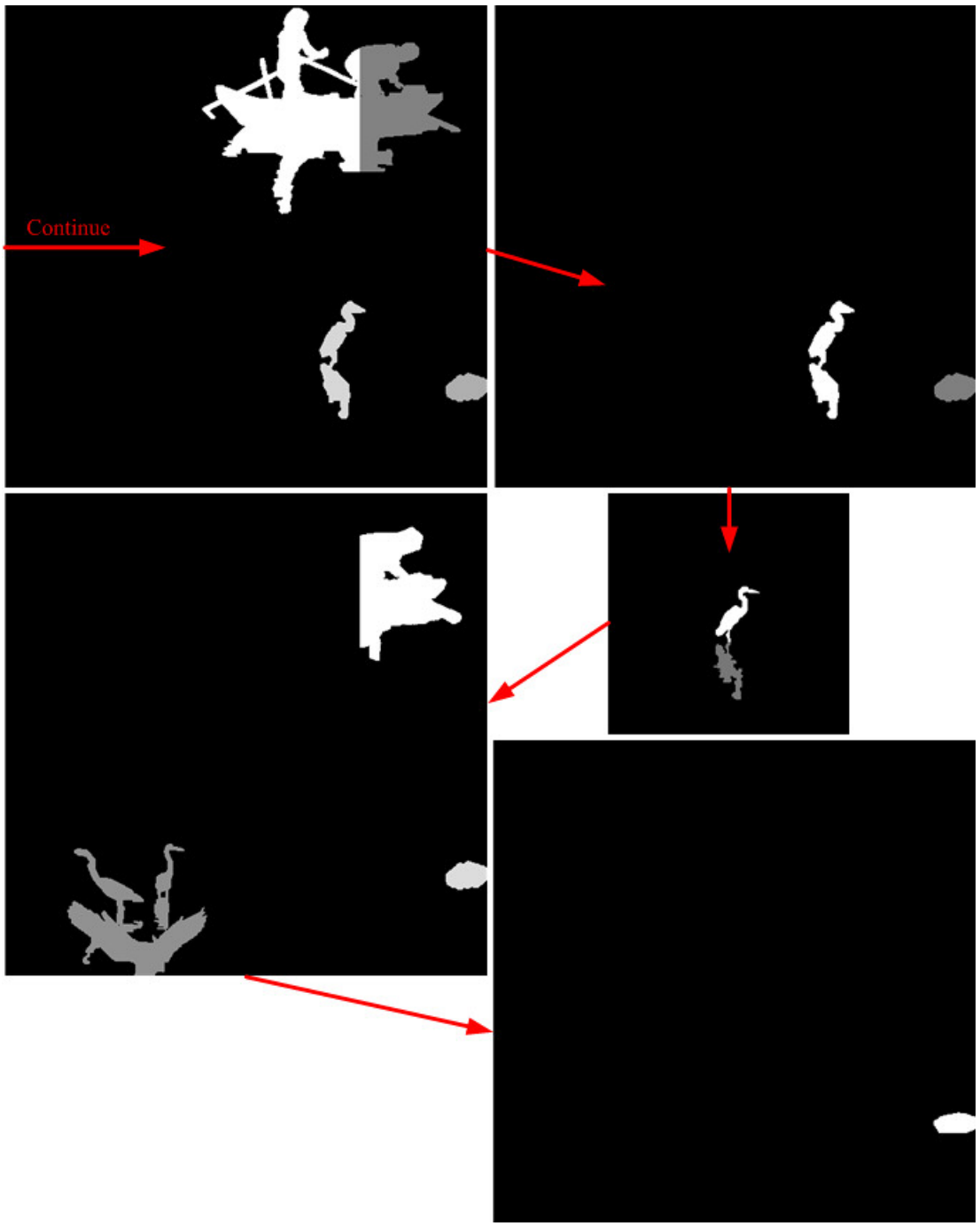

Figure 6.6: Saliency maps continued from Figure 6.5. Suppressed saliency maps due to temporary inhibition of return are clearly shown in the left column.

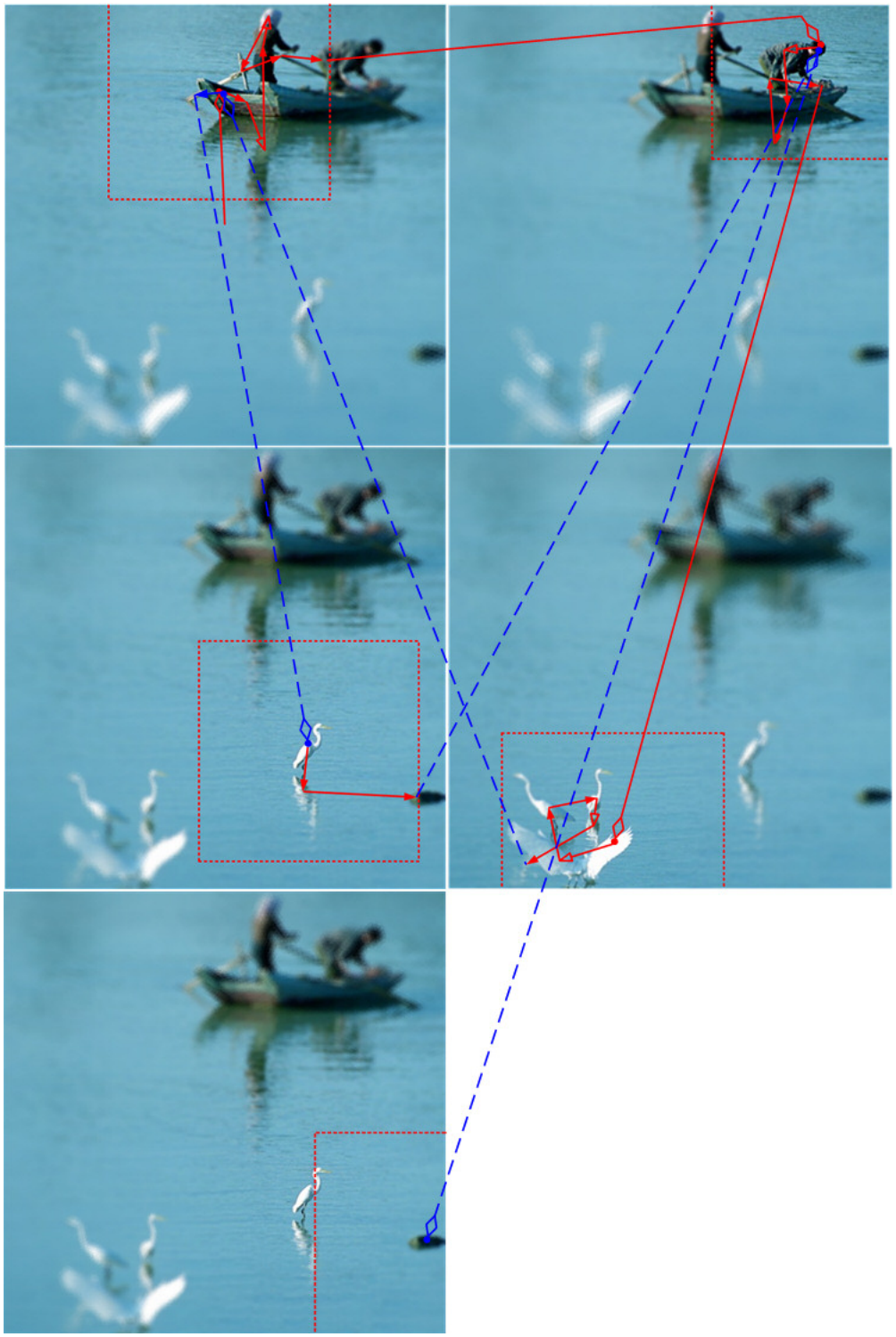

Figure 6.7: Attention-driven saccadic eye movements with attentional shifts within an attention window. Red windows: 256×256 attention windows during saccades; Red solid arrows: shifts of visual (attentional) selection; Red arrows with air diamonds: saccadic eye movements; Blue arrows with air diamonds: saccadic eye movements due to temporary inhibition of return. Blue solid arrows: attentional shifts to unattended remainder of a previously attended grouping due to temporary inhibition of return.

includes sub-groupings of a woman, a man, the boat itself, oars, their shadows reflected in the water, etc. The two people themselves are structured groupings too. The initial fixation position is assumed at the centre of the scene, i.e., position (256, 256).

The first retinal image extracted from this scene is shown at the top right image in Figure 6.4. Before the first saccade is ready to launch, the raw saliency map and the most salient grouping in this imaging are obtained from HOAM (presented in Chapter 3). Guided by the ADO mechanism, a saccade is projected to the most salient location of the most salient grouping. In consequence, the fovea is brought to a new location and a new retinal image is created following this shift. This is the top left panel in Figure 6.7, which shows the scanpath of saccades and attentional shifts. Correspondingly, the saliency map for this new foveation is re-created and adjusted over time according to the viewing change. When the fovea is fixated, the groupings within the attention window start to compete for visual attention. After several attention selections monitored by the tIOR mechanism, the saccade will jump to a new location that is outside the current attention window and wins the competition for next saccading. The previously suppressed groupings within the attention window will take part in later competitions for attention and may possibly win re-attending when their salience rises to a significant level. Figure 6.7 shows the sequence of saccadic and attentional movements, the top level grouping boat was re-attended to twice, as indicated by the blue arrows in the image at the top left of the figure. The saliency maps during saccades and (covert) attentional shifts are given in Figures 6.5 and 6.6.

It is clear that the salience of a grouping dynamically varies over time while the fovea position shifts. Its competitive capacity to gain the overt and covert attention also varies with the rise and fall of its dynamic salience even when any top-down attentional priming effects are not considered. Separate attention shifts and saccading movements are clearly shown. The human-like visual selection behaviour, i.e., primary selection by covert attention with supporting mechanism by overt saccadic eye movements, is achieved through OADS.

6.6.2 Comparison with Other Work

In order to further analyze the performance of OADS, a well-known machine vision model of space-based attention is selected, i.e., the saliency-based attention model proposed by Koch and Itti et al. [64]. Saccading behaviour is claimed to be generated by this model. (For convenience, this model is called “iModel” in the remainder of the

chapter.) A number of natural scenes randomly chosen from Itti's webpage [66] were used for the comparison.

6.6.2.1 Comparison in a Natural Scene

Figure 6.8 shows two real-world scenes and the resulting scanpaths (in red arrows) of saccades generated by iModel, at the top right and bottom left respectively. From these two scenes, we can see that some salient objects (e.g., numbers 100 and 60 in both scenes which are actually not selected by visual attention though a saccade once reaches their neighbourhood, and two close white pillars in the scene at the bottom left) were not attended and some attended objects and locations are actually not salient or nonsense for human visual attention. The weaknesses of iModel may come from the following facts:

1. iModel is only space-based without considering object-based attention. Thus the model lacks the natural strength of object-based attention and could not avoid the odd saccades in complicated real scenes. It is obvious that a larger grouping, which consists of many average salient points, could be more salient than a smaller grouping which includes a few highly salient points (see examples in the paper [133] for illustration);

2. In their model, the saccade and attention mechanisms were the same one. But psychophysical evidence shows that they are different in many ways (see Chapter 5 for the related discussion);

3. Their model did not take into account the varying resolution like the human eye's foveated imaging. Thus, the saliency mapping obtained from a scene is constant and may not reflect the saliency changes as the fovea jumps from one place to another;

4. They used constant inhibition and did not consider a time factor and saliency dynamics with eye and attention moving over time and across resolutions.

The above are also the main difference of OADS compared with other previous saccading models. The results obtained from OADS on the scene at the top left in Figure 6.8 are given in Figure 6.9 which shows the saliency mappings during saccade and attention shifts in the scene, and in Figure 6.10 which shows the scanpath of both saccading jumps and attentional shifts. It can be seen that some region groupings, which

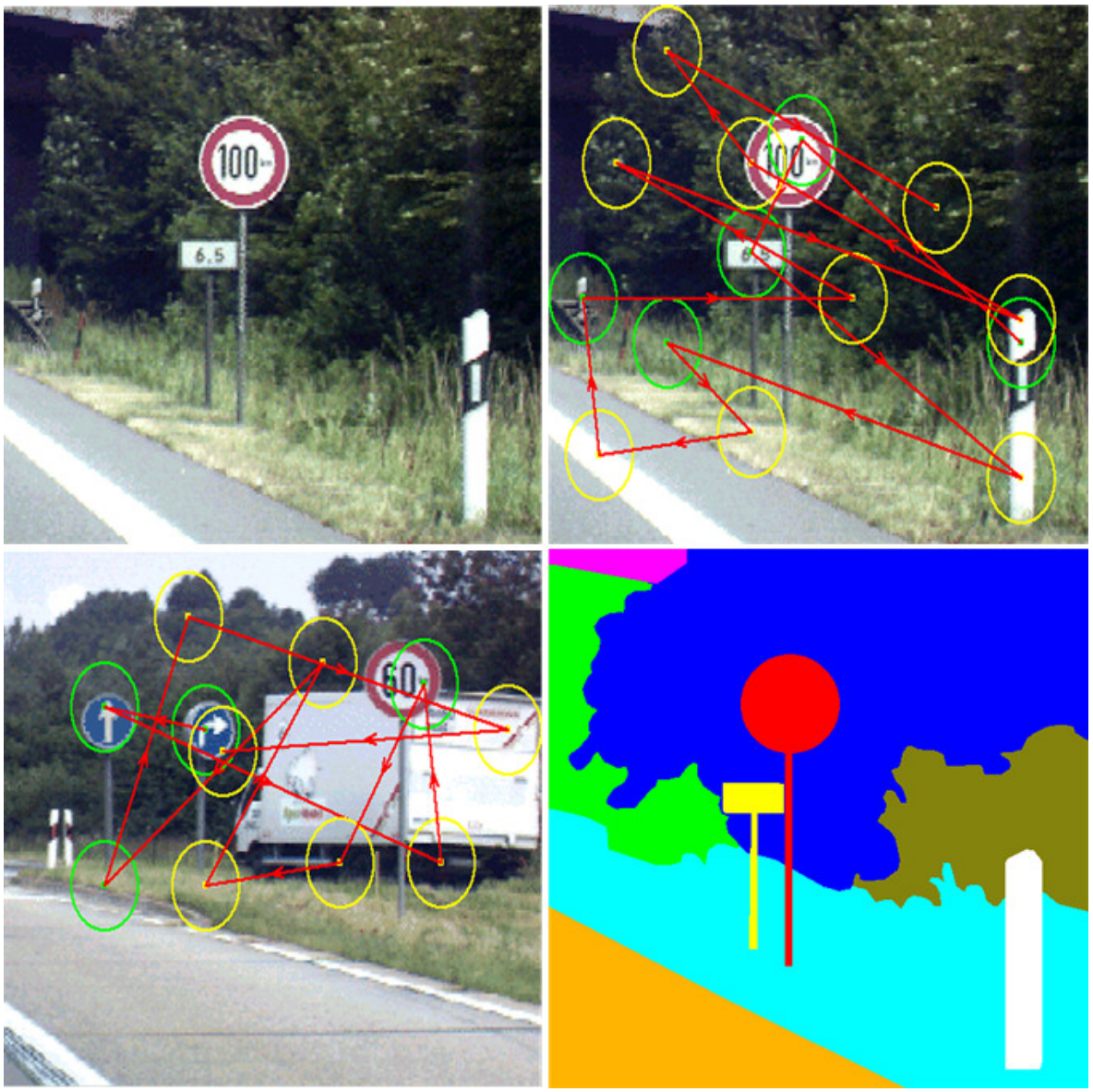

Figure 6.8: Top left: The scene (copied from Itti's homepage [66]) used for the comparison of our model with other models. Top right and bottom left: the results for two different real scenes also obtained from his homepage. Bottom right: top level groupings (shown by different colour regions) in the initial foveation when the fovea is fixated at the center of the scene.

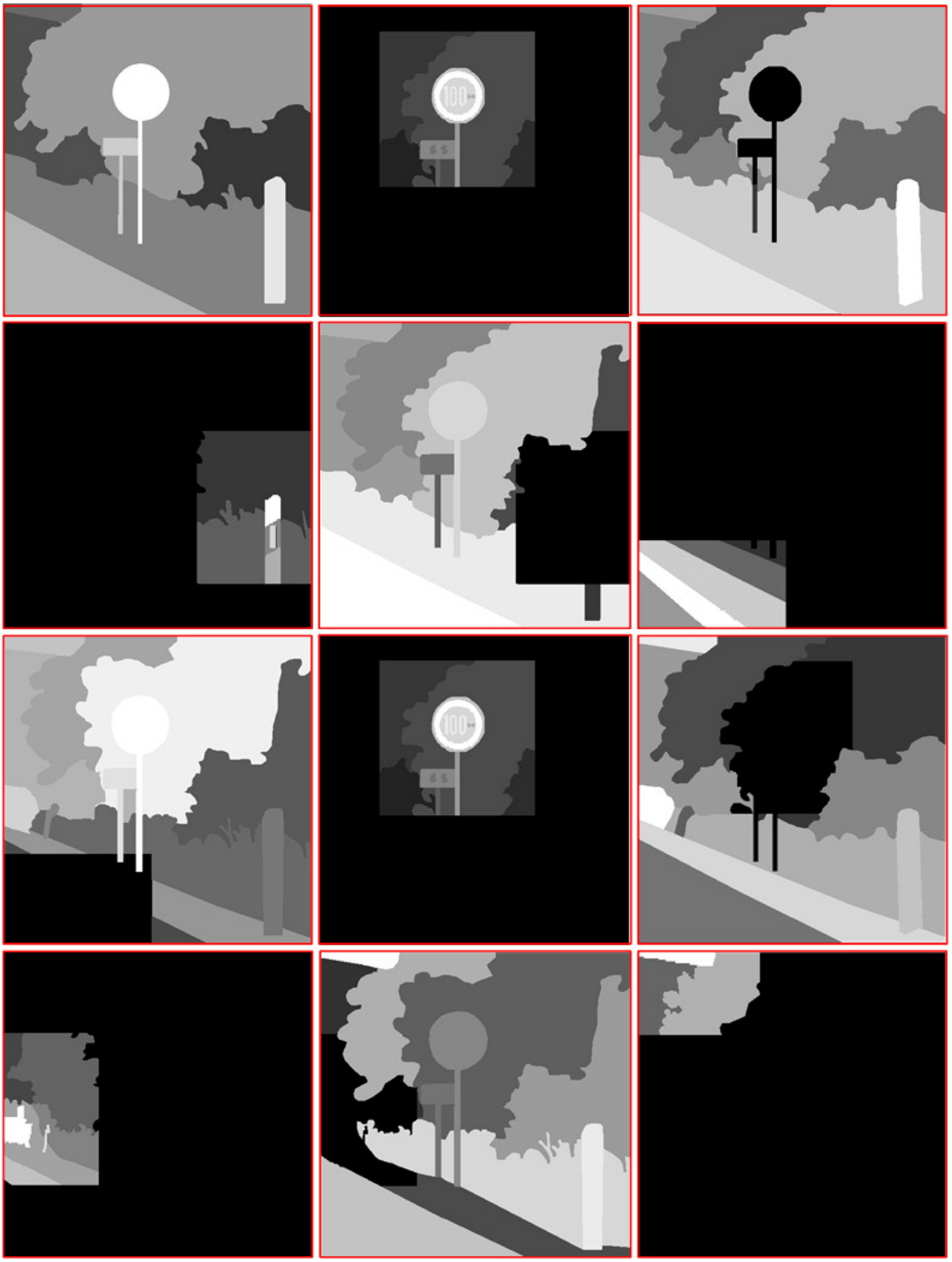

Figure 6.9: Saliency maps obtained during saccadic jumps (the order is from top left to bottom right) and attentional shifts. The dynamic variability of the groupings and their related varying saliency is clearly shown in these maps. In the figure, brighter grey scales denote higher salience magnitude and the windows denote the attentional windows. If we label the images a, b, c, d, e, f, g, h, i, j, k, l from top left to bottom right, images a,c,e,g,i,k show the globally competitive saliency mapping and the others show the saliency mapping inside the attention window.

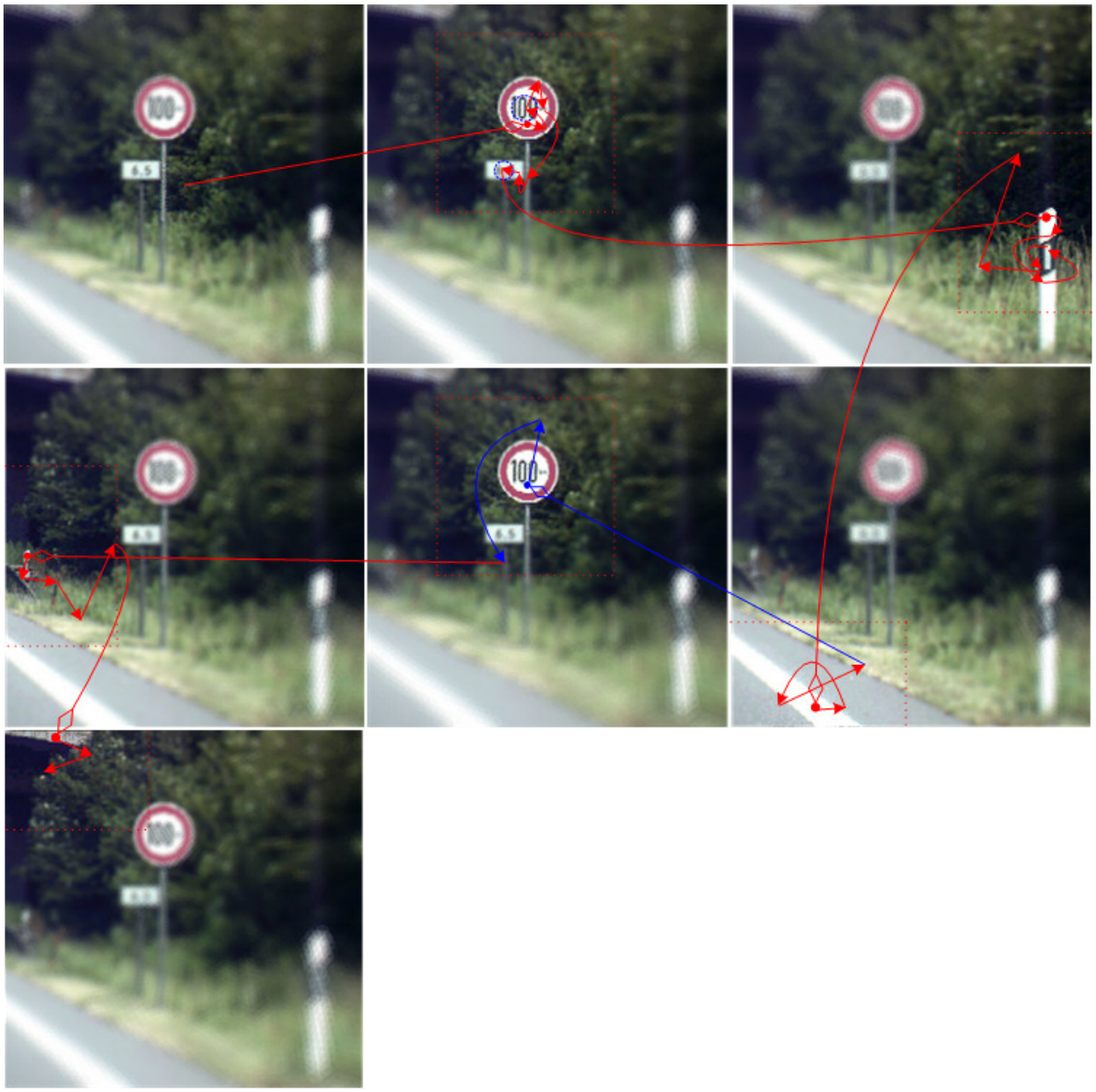

Figure 6.10: The scanpath of saccades and attentional shifts obtained from the proposed model OADS. The implications of the arrows shown here are similar to those in Figure 6.7.

Figure 6.11: Some saliency maps obtained from OADS by using only space-based computation. Top left: location-based salience map from the initial foveation. Top right: location-based saliency map from the first saccade imaging (the fovea is fixated at the biggest sign). Bottom left: location-based salience map outside the attention window when saccade is going to jump from the above fovea position. Bottom right: location-based saliency map in the attention window. Note that different Gaussian scales are used for saliency computation within and outside the attention window.

consist of almost average or unremarkably-varying salient points, such as the trees, road, etc. are attended or saccaded by population rather than individual points from one location to another. Some nonsense locations (e.g., some locations in the trees or grass) are not attended but some small highly salient objects (e.g., the number "100" or the white strip within the big white pillar at the bottom right in the scene) are attended. Notice that, in this experiment, the biggest signal plate has been saccaded twice and some objects unattended in the previous saccade around or close to the plate within the attention window have been attended, due to the temporary inhibition mechanism. The varying resolution of saccade and attention saves time and visual resources to quickly scan the objects falling at the coarse resolutions in the image periphery and scrutinize the interesting objects at the finer resolutions.

In this experiment, the behaviour of OADS shown here is clearly compatible with human visual behaviour. For some further exploration of the model performance on space-based attention/saccade, OADS also ran with a space-based saliency computation on the same scene. Some of the results are shown in Figure 6.11. Even based on this location (or pixel here) saliency computation, the salient objects attended in the above experiment are clearly shown but we can see, the orders of saccadic jumps and attentional shifts are different to the object-based attention computation due to the difference between grouping-based and location-based saliency computations.

6.6.2.2 Overall Comparison in Natural scenes

For an overall comparison with iModel, ten additional natural scenes are used. In Figures 6.12 and 6.13, the first and third rows contain the scenes and results obtained from iModel and the rows under them show the corresponding results of OADS. In each scene, OADS runs an experiment similar to that shown in Figure 6.10. But, to save space and for easy comparison with iModel, each image containing the results of OADS only shows an overall scanpath of attentional and saccadic movements. Foveated images and attention windows are omitted. To make an overall and more objective comparison between these two models, a quantitative data analysis is made. We asked ten people (three female and seven male post-graduate students and researchers with different science backgrounds including human vision research area) to count five different statistics in the ten pairs of images respectively. The five statistics are: total objects, objects selected by each model, total salient features, salient features selected by each model and redundant selection of each model. Each subject is first asked to label the total objects and salient features in the original images based on his own

Figure 6.12: Overall performance comparison of OADS with iModel. Odd rows: results obtained by iModel; Even rows: results obtained by OADS. Note that dotted circles drawn here is for improving visibility only and not denoting that attention selects these locations. Red and yellow arrows denote saccadic eye movements and attention shifts respectively. Red arrows with an open circle end indicate the first saccade.

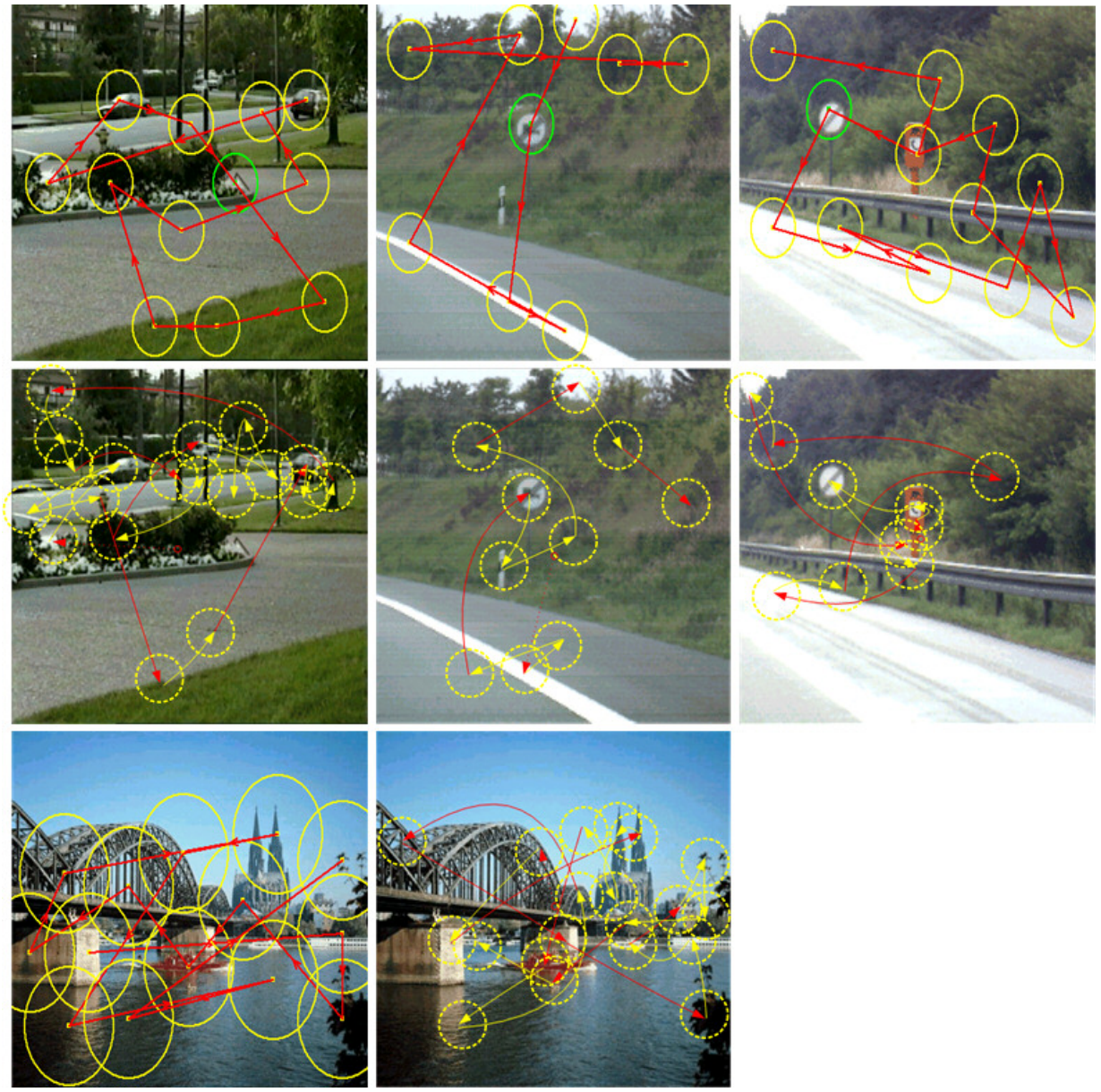

Figure 6.13: Overall performance comparison of OADS with iModel (continued slides of Figure 6.12).

Scene	Total Objects Human Selected	OADS Selected / iModel on Objects	Total Salient Features Human Selcted	OADS Selected / iModel Selected	OADS Redundant Selections / iModel Redundant Selections
1	8.3 (6.8~ 9.7)	7.9 (7.0~8.8)	24 (15.8~ 32.2)	15.9 (11.7~20)	0.3 (0~0.6)
		4.4 (3.4 ~5.5)		5.6 (4.8~6.3)	2 (1~3)
2	10.3 (9.4~11.2)	10.1 (9.4~10.9)	26.1 (19~33.3)	16 (11.7~20.3)	0.1 (0~0.4)
		5.9 (5.3~6.4)		8.3 (7.4~9.2)	2.4 (2~2.8)
3	8.1 (7.4~8.9)	8.1 (7.4~8.9)	14.6 (14.2~15)	10.3 (9.1~11.5)	0.1 (0~0.4)
		5.4 (5~5.8)		7.4 (7~7.8)	4.7 (4.2~5.3)
4	5 (4.6~5.4)	5 (4.6~5.4)	5.9 (4.8~6.9)	4.9 (3.6~6.1)	0.1 (0~0.4)
		3.9 (3.6~4.1)		2.4 (2~2.8)	10.4 (9.5~11.4)
5	12 (11.3~12.7)	12 (11.3~12.7)	17.4 (14.4~20.4)	12.9 (12.1~13.7)	0.1 (0~0.4)
		6.9 (6.1~7.7)		9.4 (8.6~10.3)	4.4 (3.6~5.3)
6	13.6 (11.9~15.2)	13 (12.1~13.9)	21.3 (20~22.5)	12.9 (12.1~13.6)	0.1 (0~0.4)
		5 (4.6~5.4)		6.4 (6~6.8)	3.4 (3~3.8)
7	21.1 (17.6~24.7)	17.4 (16.6~18.3)	17.1 (11.8~22.5)	11 (8.5~13.5)	0 (0~0)
		6.9 (5.9~7.9)		5.9 (3.5~8.3)	3.1 (2.4~3.9)
8	6.9 (5.5~8.2)	6.9 (5.5~8.2)	8.4 (6.3~10.5)	6.4 (5.6~7.3)	0.3 (0~0.6)
		4.3 (3.9~4.6)		3.6 (3.2~4)	3.6 (3.2~4)
9	7.4 (6.7~8.2)	6.1 (5.5~6.8)	9.1 (6.8~11.5)	7.1 (6.4~7.9)	0.1 (0~0.4)
		4.3 (3.9~4.6)		5 (4.3~5.7)	4.3 (3.9~4.6)
10	11.7 (9.9~13.5)	11.7 (9.9~13.5)	18.9 (12.4~25.3)	16.9 (12.2~21.6)	0.1 (0~0.4)
		6.6 (5.5~7.6)		8.4 (6.3~10.5)	3 (2.4~3.6)
Total	104.4 (99.4~109.4)	99.6 (95.9~103.3)	162.9 (149.3~176.5)	114.1 (104.9~123.4)	1.6 (0.5~2.6)
		55.3 (53.4~57.1)		62.4 (57.6~67.3)	41.3 (39.3~43.2)

Figure 6.14: Quantitive comparison of OADS with iModel. The data are produced by statically averaging the label work of ten human subjects. The main number is the mean and the bracketed values are the 95% confidence intervals.

judgement standard for image segmentation (or perceptual grouping), salient features and redundant selections after the introduction the three feature definitions that are given below. Then using the same standard, they labelled the corresponding result images obtained from the two models. The results are shown in Table 6.14 where the comparison results are averaged over ten people. The numbers in brackets are the corresponding result ranges of the ten people based on 95% confidence intervals [14].

In this table, an "object" is actually a proto-object [27, 122] and defined as a hierarchically structured grouping which is segmented from its surround. All of its components share at least one or more common properties (e.g., colour, intensity, orientation, texture, etc. or other Gestalt principles for perceptual grouping). Using this approach to (manual) segmentation, a region (e.g., a piece of sky or a lawn) may be classified as an "object". Because iModel is space-based and can only perform location-based selection, for a comparison with object-based selection implemented by OADS, the table uses "on object" for iModel to denote a shift of attention into an object. This means, once a locus of attention is within an object, it is approximately counted as an "on object". A "salient feature" is defined as a non-object but is salient in a scene, such as a salient edge, corner, blob, signal, number, character or letter, etc. A "redundant selection" (or nonsense selection) is defined as a selection of an "object" or a "salient feature" more than once while that "object" or "salient feature" is not hierarchical and has no part more salient or special than the remainder.

From the table, it is clear that iModel generally performed poor object-based selection, acceptable location-based selection, and had lots of questionable or nonsense selection. In contrast, OADS generated much better visual selection by on average $(99.6-55.3)/10 = 4.43$ more "on objects" (in total 104.4 objects in the ten scenes), $(114.1-62.4)/10 = 5.17$ more "salient feature" selection (in total 126.9 "salient features" in the ten scenes), and $(41.3-1.6)/10 = 3.97$ fewer redundant selections in each of the ten scenes. The data shows a consistent trend which is very positive to our work though the ten subjects had some different image segmentation/perceptual grouping judgements between each other. The overall better performance of our work is also reflected by the ranges in brackets. In this comparison, visual hierarchical selectivity from coarse to fine scales, foveal nonuniform resolution sensing, time-varying and resolution-varying competition and saliency mapping, temporary inhibition of return, and attention-driven saccadic eye movements have not be considered.

The better performance of our work is not surprising because fewer redundant selections and more object-based selections are the inherent results. One may criticize

that the model gains benefit from the good manual image segmentation or perceptual grouping. We accept that our work can benefit from any good segmentation approach (regardless of whether manual, semi-manual or automatic) which does not benefit the space-based attentional models. As already discussed in Section 4.5, visual perceptual grouping or segmentation is tightly linked to visual attention especially object-based attention and both of them benefit from each other. Because of this benefit, object-based attention exploits this advantage to naturally avoid many redundant or nonsense selections and achieve object-based hierarchical selection of the objects that perceptual grouping process provides. The inherent strength of the proposed model does not rely on the log-polar imaging and low-level feature extraction processing. Rather, the strength mainly comes from the grouping-based competition for visual attention. Therefore, even if the model does not use a retina-like sensor and directly works on the uniformly sampling images as many other attention models including our previous work did in Chapter 5, better results than those of the previous space-based approaches can still be anticipated.

6.7 Conclusion

This chapter presented an object-based attention-driven saccading model (OADS) to implement a two-level system of object-based attention selection with saccadic eye movements. The model also incorporates dynamic grouping-based competition and saliency mapping in a space-time context, and demonstrates its performance on complicated natural scenes. Behaviour similar to human saccadic eye movements is shown by the results obtained from the real-world scenes. By comparison with other successful models in the machine vision area, OADS shows much better and more biologically convincing visual selection behaviour in real-world visual environments. OADS is outstanding itself by the following properties: spatio-temporal dynamic grouping-based saliency mapping and competition for attention/saccade, human-like foveation imaging, temporary inhibition of return, separate processing levels for covert attention and overt eye movements, and object-based attention-driven saccadic eye movements. These advantages appear to endow OADS with visual attention as the primary visual selection mechanism and saccadic eye movements as a supporting role – a feature of human visual attention system and better performance than previous machine vision attention/saccading models.

Chapter 7

Conclusions

Human vision uses visual attention to select interesting information and employs attention-guided saccadic eye movements to explore visual environment for further scrutiny of a scene. Previous machine vision systems have not yet exploited object-based attention and two-level covert attentional-overt saccading shifts integrated in one selection framework. This results in non biologically-plausible visual behaviour and worse performance in these systems, especially when they deal with complex visual selection in real-world visual environment.

The work presented in this thesis, in contrast, develops a Hierarchical Object-based Attention Framework (HOAF) to provide machine vision with human-like, effective visual selection behaviour. HOAF adopts object-based attention and uses grouping-based competition to implement object-based hierarchical selectivity and to integrate object-based and space-based attention. Visual attention and saccadic eye movements are built into distinct processes and work together to accomplish complex visual selection tasks in a spatio-temporal context. HOAF employs a log-polar retina-like sensor to nonuniformly sample the field of view. In the meantime, object-based attention undertakes the inspection of interesting "proto-objects" (or groupings) in the attended area with higher resolution surrounding the fovea, while potentially interesting objects in the coarse resolution periphery of the field of view are surveyed with the help of attention-guided saccadic eye movements. The overlapping retinal images resulting from gaze shifts over time are effectively dealt with by mapping them into the unified coherent representation – spatio-temporal grouping saliency mapping. The competition for visual attention/saccade is supervised by the common mechanisms of competition pool of attention and temporary inhibition of return working in a spatio-temporal context. Similar to human visual behaviour, complicated hierarchical objects

can be re-analyzed more than once in HOAF. Tested on a number of synthetic images and real-world natural scenes as well as compared with other previous famous work, HOAF demonstrates better visual attention and saccadic eye movements performance which is showed to concur with the main findings found in psychophysical research on object and space-based visual attention. This is the first time that object-based attention modeling with distinct visual covert and overt selection is implemented in a machine vision system.

The rest of this chapter summarises the novel contributions of the Hierarchical Object-based Attention Framework (HOAF) and then discusses some useful directions for extending the work.

7.1 Contributions

Our object-based attention framework HOAF distinguishes itself from previous machine vision studies on visual attention and saccadic eye movements by a number of novelties:

Foundations for Computational Object-Based Visual Attention

HOAF is the first computable object-based attention framework for machine vision. Two primary foundations are provided for modeling visual attention in machine vision: (1) a general hierarchical object-based attention framework integrating space-based attention; (2) a human-like visual selection system with (covert) attention as the primary selection mechanism and overt saccadic eye movements as a supporting role. In addition, to make HOAF more general and applicable, novel mechanisms for spatio-temporal grouping-based saliency mapping and competition, object-based hierarchical selectivity, and temporary inhibition of return are developed.

Object-Based and Space-Based Attention Integration

The integration of object-based and space-based attention has not been explored in previous machine vision systems and is first implemented in our object-based attention framework HOAF. Object-based attention is implemented through the approach of grouping-based saliency mapping and competition. By this grouping-based approach, space-based attention is naturally integrated. Visual attention operates at multiple com-

petitive levels for selectivity by objects, features and their structured groups based on their common underlying units of visual selection – groupings. Object-based hierarchical selectivity is therefore achieved.

Hierarchical Selectivity

Hierarchical selectivity is a natural behaviour held by our object-based attention framework HOAF. With the help of hierarchical selectivity, HOAF is the first machine vision system that can effectively perform hierarchical object-based selection for features, objects, regions, and their groupings from coarse to fine resolution. Experimental results with synthetic images demonstrate HOAF possesses visual selection behaviour compatible with the main findings in psychophysical research on visual attention. Through experiments with real-world natural scenes, HOAF also shows improved ability to undertake complex object-based attentional selection.

Grouping-Based Spatio-Temporal Saliency Mapping and Competition

Grouping-based saliency mapping and competition are the key processes of our object-based attention framework HOAF. Saliency mapping has been broadly employed by previous machine vision systems to model visual attention. However, these attention systems used location-based saliency mapping to evaluate the competition for attention in a visual environment. This does not reflect the actual competition which is between objects rather than between individual or location-like features of the image. In contrast to previous work, saliency evaluation is achieved in HOAF by an original approach using grouping-based saliency mapping which is designed to measure how dominant an object/a grouping is when it competes with other objects/groupings in a spatio-temporal context for attention. That is, grouping saliency mapping and competition are dynamic and vary with space-time. This spatio-temporal property derives from the following facts:

- The salience of a grouping is evaluated by integrating contrasts from its local surround at small spatial scales to its global surround at large spatial scales.

- The structure and surround of a grouping vary with different spatial resolutions resulting in dynamic saliency mapping and competition.
- The grouping saliency mapping of a field of view varies over time due to the time-varying inhibition of return and multiple nonuniform resolution sensing by foveal movements over time. Correspondingly, the competition between groupings for attention varies over time.

The spatio-temporal dynamics of grouping saliency mapping and competition provides the spatio-temporal hierarchical selectivity behaviour needed for visual attention and saccadic eye movements.

Distinct Attention Selection and Attention-Driven Saccadic Eye Movements

The human visual system uses (covert) attention to achieve visual selection and employs attention-guided saccadic eye movements to extend this visual selectivity when exploring large-scale visual environments. Our Hierarchical Object-based Attention Framework HOAF is the first machine vision system to combine these two primary visual mechanisms into one integrated selection system but makes a distinction between them. In this framework HOAF, where a saccade goes to next and when it starts to jump are determined by the competition for attention. Visual (covert) selection by attention and overt jump orienting by saccades use different mechanisms and work together for the coherent visual selection. These two kinds of covert and overt shifts are clearly modelled at two levels. Similar to human visual selection behaviour, the proposed object-based attention framework HOAF uses attention to scrutinize interesting objects around the fovea and makes use of attention-guided saccadic eye movements to extend visual attentional selection in the field of view. This important feature of distinct shifts due to (covert) attentional selection and saccades in one selection system concurs with one of the main features of human attention system but was not explored in machine vision. HOAF clearly shows this kind of visual behaviour of distinct attentional shifts and attention-guided saccading movements on real-world natural scenes. Also, during the movements of attention and saccading, previously attended objects can be reattended/refixated more than once in a spatio-temporal context due to temporary Inhibition Of Return (tIOR).

Psychophysical-Plausibility and High Effectiveness

The work presented in this thesis is inspired by the recent major achievements of psychophysical and neurobiological research on visual attention and saccadic eye movements. Each of the original mechanisms proposed in the work has a strong background of biological and psychophysical plausibility. This thesis demonstrated visual behaviour of our Hierarchical Object-based Attention Framework (HOAF) compatible with the main findings in the above areas. Moreover, HOAF achieved much better performance and higher effectiveness when compared with other well-known research on a number of real-world natural scenes.

7.2 Future Work

The work reported in this thesis is important because it provides machine vision the first successful modeling of object-based attention and attention-guided saccadic eye movements. However, there are many improvements and extensions that could be made to the work. Some of them have been addressed at the end of the relevant chapters. Other interesting areas of the research are suggested below.

Automatic Grouping

Perceptual grouping or object-based visual segmentation is very important because objects are the underlying units of visual attentional selection. But, what is the relationship between segmentation and attention? If viewed from the conventional preattentive/attentive dichotomy, a question is raised naturally: can an object hierarchy be found and objects be segmented preattentively? Recent studies corrected this simple serial two-stage processing assumption and revealed that human vision incorporates multiple levels of processing while attention is best regarded as an emergent state or "umbrella-term" for multiple selective processes rather than a single process. Furthermore, these studies suggested mutual constraints between segmentation and attention. "Many forms of attentional selectivity and grouping are implemented in the brain by interactions between multiple levels of processing" (see [27] for a review of these relevant studies).

Therefore, an effective and biologically-plausible approach for automatic grouping should consider both attention and segmentation together. An ideal solution should further involve perceptual experience-based knowledge, reasoning, and learning.

Although many segmentation approaches have been developed in the literature, most of them only work in a limited range of visual environments and many of them require manual help to obtain good results in general applications. In addition, most image segmentation approaches do not consider the hierarchical structures in the real visual world. Perceptual grouping is therefore usually difficult to obtain. This is because general segmentation and perceptual grouping are context-based as well as experience-based and involve complex visual or nonvisual reasoning from top-down interactions so that it is difficult to build a general framework.

As suggested above, a possible approach to building a more general automatic grouping system may consider a neural-based architecture integrating both bottom-up and top-down processing based on grouping competition. This kind of neural architecture may consist of a low-level feature extraction network, a contrast-based grouping network and a knowledge network.

The low-level feature information extracted from the input images by context-sensitive receptive fields of neurons is used to produce feature contrast and similarity input to the grouping network for the classification and contour processing (e.g, Grossberg et al. work [54]).

The knowledge network may use some kinds of object templates based on several Gestalt perceptual grouping rules to generate the top-down matching input to the grouping network. Then, locally short-range competitive and globally long-range cooperative interactions in the grouping network are activated by the bottom-up input and feedback matching input from the top-down knowledge network. During the interactions, visual attention which is taken as an emergent state plays the visual selection role to bias or enhance the activations favouring the matching between the bottom-up and top-down interactions. The unfavourable activations will be suppressed at the same time.

Consequently, the neurons with enhanced activations will tend to form groupings through synchronization. The hierarchical structure of groupings can be achieved through the excitation and inhibition links within and between layers of the grouping network.

The above approach can achieve a more general usage of visual segmentation/perceptual grouping in normal visual environments because the experience-based knowledge, reasoning and learning can be incorporated in the knowledge network to improve the performance and visual attention can be used to greatly reduce the computational complexity when dealing with intricate visual scenes.

General Top-down Attentional Priming

The Hierarchical Object-based Attention Framework (HOAF) presented in this thesis suggested a simple mechanism for top-down attentional priming and has not incorporated knowledge-based such as object-based template interaction with attention. But the top-down modulatory influence on attention has not been completely explored. Also, it does not have a general top-down priming architecture.

However, it is now well known that top-down attentional priming greatly affects visual attention. In other words, visual attention is modulated by both bottom-up and top-down influences. Because top-down influence involves complex processes such as visual and nonvisual reasoning, perceptual experience, and learning, etc., building a general and reasonable top-down control structure is a very challenging but valuable research area. It will also benefits object recognition or perceptual grouping.

A possible approach is to use knowledge-based neural networks that encode the general proto-object templates built upon Gestalt perceptual grouping rules and specific information relevant to the current visual tasks. Learning can also be incorporated to benefit the knowledge network to obtain more experience for adapting to the more general applications. Biasing information or attentional priming will result from the competition between the different kinds of templates through the interactions with bottom-up context-sensitive input.

Visual Selection for Motion and 3D Objects

The work presented here has not incorporated motion and 3D information and the experiments therefore all use static 2D visual scenes, although in theory the work can be applied to video sequences for the visual salient object selection task which does need object recognition since each frame of the videos is static.

Similarly, if visual tasks do not consider object recognition or smooth pursuit, our approach in theory can be easily extended to work with real motor-driven cameras to implement object-based visual selection with saccadic eye movements. But this will require perceptual grouping available in advance. However, as we showed in Chapter 4, the proposed work can also be used to achieve location-based visual selection (as in the space-based attention models) without employing perceptual grouping.

Motor-driven camera systems can take advantage of the attention-driven mecha-

nism to guide their fixation shifts and make use of the hierarchical object-based attention mechanism to achieve visual hierarchical selectivity from coarse to fine scales when the fovea is fixated for a certain time.

But for general usages in machine vision, many visual tasks also require motion and 3D-based application. Therefore, extending the current work to a 3D object-based visual selection framework in dynamic scenes is a useful direction for future research. Taking motion and other 3D features into account, this extended selection framework could perform smooth pursuit and vergence eye movements.

Inhibition of Return and Objects-Based Attention

Inhibition of return is known to be coupled with visual attention activities (see the related discussion in Section 6.4). An interesting question arising from our current work is: "In what way does inhibition of return function on object-based visual selection?" This may be not a problem for space-based attention models because they only need to inhibit an attended location and its neighbourhood but it would be a problem to the space-based attention models that incorporate region-based (or blob-based) spatial selection because an inhibition of a selected region is required.

Two possibilities are involved in the above question: 1) Is an attended object transiently inhibited by equally suppressing the saliency of its whole structure or all members at the same time or 2) by first starting the suppressing effect from a (some) suppressing center(s) within it and then spreading the inhibition to the whole object over time? If the former hypothesis is true in human visual attention system, it will be easily simulated in the algorithm. Otherwise, the mechanism of inhibition of return may not straightforward. One of the factors that may be involved is the diffused manner of inhibition. How this spreading procedure happens and does it involve distance effects because parts of an object have different distances from the suppressed center(s)? Another factor is where should the suppressed center(s) be, at center of the mass of the attended object or elsewhere? In the psychophysical attention literature, we have not found satisfactory findings to answer either of these questions. This may be because the research on object-based attention is very young and many related questions are still open.

In our work, we employed both mechanisms. The first one is used in the Hierarchical Object-based Attention Model (HOAM) which simply suppresses the entire attended grouping or sub-grouping uniformly at a time. The second one is used in

the Object-based Attention-Driven Saccading model (OADS) that takes each component of an attended object as a suppression center to suppress the surround within that object so that components may be suppressed nonuniformly. Consequently, their saliency rises nonuniformly, too. But at the moment of inhibition, these two methods produce a similar effect for the inhibition of return so that visual attention can shift to the next object. Further investigation to obtain a biologically-plausible inhibition of return mechanism for object-based attention will be very interesting and useful to both psychophysical and machine vision research.

Appendix A

Glossary

Attentive Processing refers to attentional processing.

Covert Internal and unobservable processes that are used to refer to (or emphasize) visual attentional shifts that switch between different visual selections and are separable from eye movements.

Deployment Unit or underlying unit refers to an object or a location which is selected by visual object-based or space-based attention.

Grouping is defined as a hierarchical unit or structure which is segmented from its surround based on Gestalt perceptual grouping (or segmentation) rules. In this definition, a grouping may be a structured "proto-object" but also includes a segmented hierarchy of spatial regions in a visual scene.

Object-Based theories of attention propose that attention selects an "object" rather than a location of space on which additional processing resources are then focused. The "object" term used here to describe the underlying unit of attentional selection is best thought of as a (hierarchical) "proto-object" (or a grouping) derived from segmentation processes rather than a solid-like object we experienced in normal life (e.g., a apple).

Overt External and observable processes that are used to refer to (or emphasize) eye movements that help visual attention to gain access to interesting information in the periphery of the field view for higher resolution processing of visual selection.

Preattentive A traditional term that refers to the process that occurs before the attentional operation.

Selective Levels or multiple selection or hierarchical selectivity refer to the flexible allocation of attentional resources to different processing levels or subsets of visual information, including filtering by location or other properties, object or group of objects.

Space-Based theories of attention assume that attention selects a location (or a region) of space rather than treating a location as a feature of an object as object-based attention theories hold.

Visual Attention is a mental process referring to the selective aspects of visual perception that enables an observer to recruit greater resources for processing interesting information of visual environment more fully than the unselected others.

Appendix B

List of acronyms

ADO Attention-Driven Orienting mechanism which is proposed to guide saccadic eye movements based on the competition between the groupings within and the groupings outside the attention window.

HOAF Hierarchical Object-Based Attention Framework.

HOAM Hierarchical Object-Based Attention Model which is the foundation of HOAF.

OADS Object-Based Attention-Driven Saccading model which is built upon HOAM to achieve visual attentional selection with attention-guided saccadic eye movements in distinct levels in HOAF.

tIOR Temporary Inhibition of Return which is proposed to transiently suppress an attended grouping or sub-grouping to prevent attention from immediately returning so as to force attention shifts. After attention shifts, a temporarily suppressed grouping starts to update its saliency over time so that it may possibly gain visual attention again at other time.

Appendix C

List of Parameters

θ: preferred orientation in Gabor filters.

$\theta \in [0^0, 45^0, 90^0, 135^0]$ or $[0^0, 22.5^0, 45^0, 67.5^0, 90^0, 112.5^0, 135^0, 157.5^0]$

α, β: weighting coefficients in Eq. 4.12. $\alpha = \beta = 1$

$\hat{n}$: maximum of the width and length of the feature maps at the current computing level of pyramids.

ρ: integer parameter of Gaussian weighting function in Eg. 4.13 is set to be from 2 to 50 in different experiments.

γ_{CI}, γ_O: weighting coefficients in Eq. 4.21. $\gamma_{CI} = \gamma_O = 1$

α: constant in Eq. 6.4 and is set to be 0.5 in the experiments.

β: constant in Eq. 6.5 and is set to be 1.

Bibliography

[1] C. H. Andersen and D. C. Van Essen, "Shifter circuits: a computational strategy for dynamic aspects of visual processing," *Proc. Natl. Acad. Sci.*, USA, 84, pp. 6297-6301, 1987.

[2] I. Ahrns and H. Neumann, "Space-variant dynamic neural fields for visual attention," *Proc. IEEE Computer Vision and Pattern Recognition*, Fort Collins, CO., pp. 313-318, 1999.

[3] G. Backer, B. Mertsching and M. Bollmann, "Data- and model-driven gaze control for an active vision system," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(12), pp. 1415-1429, 2001.

[4] S. Baluja and D. Pomerleau, "Dynamic relevance: Vision-based focus of attention using artificial neural networks," *Artificial Intelligence*, 97, pp. 381-395, 1997.

[5] S. Baluja and D. Pomerleau, "Expectation-based selective attention for visual monitoring and control of a robot vehicle," *Robotics and Autonomous Systems*, 22, pp. 329-344, 1997.

[6] M. Behrmann, R. S. Zemel and M. C. Mozer, "Occlusion, symmetry and object-based attention: reply to Saiki (2000)," *Journal of Experimental Psychology: Human Perception and Performance*, 26(4), pp. 1497-1505, 2000.

[7] G. Bonmassar, E. L. Schwartz, "Space-Variant Fourier Analysis: The Exponential Chirp Transform," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(10), pp. 1080-1089, 1997.

[8] D. E. Broadbent, *Perception and Communication*, London: Pergamon Press, 1958.

[9] C. Bundesen, "Cognitive psychology," In A. Kramer, G. H. Cole and G. D. Logan (Eds), *Converging Operations in the Study of Visual Selective Attention*, pp. 1-44, Washington, DC: American Psychologuea Association, 1996.

[10] C. Bundesen, "A computational theory of visual attention," *Phil. Trans. R. Soc. Lond. B*, 353, pp. 1271-1281, 1998.

[11] P. Burt, "Attention mechanisms for vision in a dynamic world," In: *Proceedings Ninth International Conference on Pattern Recognition*, Beijing, China, pp. 977-987, 1988.

[12] R. H. S. Carpenter, *Movements of the Eyes*, London: Pion, 1988.

[13] G. Carpenter, S. Grossberg and G. Lesher, "The representation of visual salience in monkey parietal cortex," *Nature*, 391, pp. 481-484, 1998.

[14] P. R. Cohen, *Empirical Methods for Artificial Intelligence*, Cambridge, MA. MIT Press, 1995.

[15] E. J. Chichilnisky and B. A. Wandell, "Trichromatic opponent color classification," *Vision Research*, 39(20), pp. 3444-58, 1999.

[16] J. J. Clark and N. Ferrier, "Modal control of an attention vision system," *Proc. IEEE Inter. Conf. Computer Vision*, Tarpon Springs, FL., pp. 514-523, 1988.

[17] J. J. Clark, "Spatial attention and latencies of saccadic eye movements," *Vision Research*, 39(3), pp. 583-600, 1998.

[18] C. L. Colby and M. E. Goldberg, "Space and attention in parietal cortex," *Annu. Rev. Neurosci.*, 22, pp. 319-49, 1999.

[19] V. Concepcion and H. Wechesler, "Detection and localization of objects in time-varying imagery using attention, representation and memory pyramids," *Pattern Recognition*, 29(9), pp. 1543-1557, 1996.

[20] W. Cowan, "Evolving conceptions of memory storage, selective attention and their mutual constraints within the human information-processing system," *Psychol. Bull.*, 104, pp. 163-191, 1988.

[21] F. Crick and C. Koch, "Towards a neurobiological theory of consciousness," *Seminars in the Neurosciences*, 2, pp. 263-275, 1990.

[22] http://www.dai.ed.ac.uk/CVonline, June 2003.

[23] R. Desimone and J. Duncan, "Neural mechanisms of selective visual attention," *Ann. Rev. Neurosci.*, 18, pp. 193-222, 1995.

[24] R. Desimone, "Visual attention mediated by biased competition in extrastriate visual cortex," *Phil. Trans. R. Soc. Lond. B*, 353, pp. 1245-1255, 1998.

[25] J. Dias, H. Araújo, C. Paredes and J. Batista, "Optical normal flow estimation on log-polar images. A solution for real-time binocular vision," *Real-Time Imaging*, 3, pp. 213-228, 1997.

[26] J. Driver and G. C. Baylis, "Attention and visual object segmentation," In R. Parasuraman (Ed.), *The Attentive Brain*, pp. 299-25, Cambridge, MA: MIT Press, 1998.

[27] J. Driver, G. Davis, C. Russell, M. Turatto and E. Freeman, "Segmentation, attention and phenomenal visual objects," *Cognition*, 80, pp. 61-95, 2001.

[28] J. Duncan, "Selective attention and the organization of visual information," *J. Exp. Psychol.*, 113, pp. 501-517, 1984.

[29] J. Duncan and G. W. Humphreys, "Visual search and stimulus similarity," *Psychological Review*, 96, pp. 433-458, 1989.

[30] J. Duncan, "Target and non-target grouping in visual search," *Perception and Psychophysics*, 57(1), pp. 117-120, 1995.

[31] J. Duncan, "Cooperating brain systems in selective perception and action," In: T. Iaui and J.L. McClelland (Eds.), *Attention and Performance XVI*, Cambridge, MA: MIT Press, pp. 549-578, 1996.

[32] J. Duncan, et al. "Integrated mechanisms of selective attention," *Curr. Opin. Biol.*, 7, pp. 255-261, 1997.

[33] J. Duncan, "Converging levels of analysis in the cognitive neuroscience of visual attention," *Phil. Trans. R. Soc. Lond. B.*, 353, pp. 1307-1317, 1998.

[34] H. E. Egeth and S. Yantis, "Visual attention: control, representation and time course," *Annu. Rev. Psychol.*, 48, pp. 269-97, 1997.

[35] R. Egly, J. Driver and R. Rafal, "Shifting visual attention between object and locations: Evidence from normal and parietal lesion subjects," *J. Exp. Psychol. Hum. Percept*, 123, pp. 161-177, 1994.

[36] S. Engle, X. Zhang and B. A. Wandell, "Colour tuning in human visual cortex measured with functional magnetic resonance imageing," *Nature*, 388(6637), pp. 68-71, 1997.

[37] C. W. Eriksen and Y. Y. Yeh, "Allocation of attention in the visual field," *J. Experimental Psychology: Human Perception amd Performance*, 11(5), pp. 583-597, 1985.

[38] C. W. Eriksen and J. D. St. James, "Visual attention within and around the field of focal attention: a zoom lens model," *Perception and psychophysics*, 40(4), pp. 225-240, 1986.

[39] C. W. Eriksen, "Attentional search of the visual field," In D. Brogan (Ed.), *Visual Search*, London: Taylor Francis, pp. 221-40, 1990.

[40] C. J. Erkelens and H. Collewijn, "Control of vergence: gating among disparity inputs by voluntary target selection," *Experimental Brain Research*, 87, pp. 671-678, 1991.

[41] S. Exel and L. Pessoa, "Attention visual recognition," *International Conference on Pattern Recognition*, Brisbane, Australia, 1998.

[42] M. J. Farah, M.A. Wallace and S. P. Vecera, ""What" and "where" in visual attention: evidence from the neglect syndrome," In I. A. Robertson and J. C. Marshall (Eds.), *Unilateral Neglect: Clinical and Experimental*, Hove, UK: Erlbaum, pp. 123-137, 1993.

[43] V. Ferrara and S. Lisberger, "Attention and target selection for smooth pursuit eye movements," *J. Neurosci.*, 15(11), pp. 7472-7484, 1995.

[44] G. R. Fink, R. J. Dolan, P. W. Halligan, J. C. Marshall and C. D. Frith, "Space-based and object-based visual attention: shared and specific neural domains," *Brain*, 120, pp. 2013-2028, 1997.

[45] B. Fischl, E. L. Schwartz and M. A Cohen, "The local structure of space-variant images," *Neural Networks*, 10(5), pp. 815-831, 1997.

[46] C. H. Folk, W. R. Remington and J. H. Wright, "The structure of attentional control: contingent attentional capture by apparent motion, abrupt onset and color," *Journal of Experimental Psychology: Human Perception and Performance*, 20(2), pp. 317-329, 1994.

[47] H. M. Gomes, R. B. Fisher, "Primal-sketch feature extraction from a log-polar representation," *Pattern Recognition Letters*, 24(7), pp. 983-992, April 2003.

[48] H. M. Gomes, *Model Learning in Iconic Vision*, PhD Thesis, School of Informatics, The University of Edinburgh, May, 2002.

[49] J. P. Gottlieb, M. Kusunoki and M. E. Goldberg, "The representation of visual salience in monkey parietal cortex," *Nature*, 391(6666), pp. 481-484, 1998.

[50] H. Greenspan, S. Belongie, R. Goodman, P. Persona, S. Rakshit and C. H. Anderson, "Overcomplete steerable pyramid filters and rotation invariance," *Proceedings IEEE Computer Vision and Pattern Recognition*, pp. 222-228, Seattle, Washington, 1994.

[51] W. E. L. Grimson, A. Lakshmi Ratan, P. A. O'Donnel and G. Klanderman, "An active visual attention system to 'play Where's Waldo'," *Proceedings Conference on Computer Vision and Pattern Recognition*, Seattle, WA, pp. 85-90, 1994.

[52] S. Grossberg, E. Mingolla, W. Ross, "A neural theory of attentive visual search: interactions of boundary, surface, spatial and object representations," *Psychological Review*, 101, pp. 470-489, 1994.

[53] S. Grossberg, "How does the cerebral cortex work? Learning, attention and grouping by the laminar circuits of visual cortex," *Spatial Vision*, 12(2), pp. 13-185, 1999.

[54] S. Grossberg and E. Mingolla, "Neural dynamics of perceptual grouping: textures, boundaries, and emergent segmentations," *Perception and Psychophysics*, 38, pp. 141-171, 1985.

[55] T. D. Grove and R. B. Fisher, "Attention in iconic object matching," *Proceedings British Machine Vision Conference*, Edinburgh, pp. 293-302, 1996.

[56] D. Heinke and G. W. Humphreys, "SAIM: A model of visual attention and neglect," *Proceedings International Conference on Artificial Neural Networks*, pp. 913-918, New York, NY, 1997.

[57] D. Heinke and G. W. Humphreys, "Computational Models of Visual Selective Attention: A review," In G. Houghton (Ed.), *Connectionist Models in Psychology*, Psychology Press, in press.

[58] J. M. Henderson, "Visual attention and eye movement control during reaching and scene perception," In K.Rayner (Ed.), *Eye Movements and Visual Cognition*, Springer-Verlag, pp. 260-283, 1992.

[59] J. E. Hoffman and S. Mueller, "An in-depth lock at attention," presented at the *35th Annual Meeting of the Psychonomic Society*, St. Louis, Mo., November, 1994.

[60] J. E. Hoffman, "Visual attention and eye movements," In H. Pashler (Ed.), *Attention*, Psychology Press, pp. 119-154, 1998.

[61] T. K. Horiuchi, T. G. Morris, C. Koch and S. P. DeWeerth, "Analog VLSI Circuits for attention-based, visual tracking," *The Neural Information Processing Conference*, Denver CO, pp. 706-712, December, 1996.

[62] G. W. Humphreys, "SEarch via recursive rejection (SERR): A connectionist model of visual search," *Cognitive Psychology*, 25, pp. 43-110, 1993.

[63] G. W. Humphreys, "Neural representation of objects in space: a dual coding account," *Phil. Trans. R. Soc. Lond. B*, 353, pp. 1341-1351, 1998.

[64] L. Itti, C. Koch and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11), pp. 1254-1259, 1998.

[65] L. Itti and C. Koch, "A saliency-based search mechanism for overt and covert shifts of visual attention," *Vision Research*, 40, pp. 1489-1506, 2000.

[66] http://ilab.usc.edu/bu/javaDemo/index.html.

[67] W. James, *The Principles of Psychology*, 1, New York: Dover, 1890.

[68] E. N. Johnson, M. J. Hawken and R. Shapley, "The spatial transfformation of color in the primary visual cortex of the macaque monkey," *Nature*, 4(4), pp. 409-416, 2001.

[69] F. Jurie, "A new log-polar mapping for space variant imaging. Application to face detection and tracking," *Pattern Recognition*, 32, pp. 865-875, 1999.

[70] D. Kahneman and A. Henik, "Perceptual organization and attention," In M. Kubovy and J. R. Pomerantz (Eds.), *Perceptual Organization*, pp. 181-211, Hillsdale, NJ: Erdbaum, 1984.

[71] P. Kaiser and R. M. Boynton, *Human Color Vision*, Second Edition, Published by the Optical Society of America, 1996.

[72] S. Kastner and L. G. Ungerleider, "Mechanisms of visual attention in the human cortex," *Annual Review of Neuroscience*, 23, pp. 315-341, 2000.

[73] Y. B. Kazanovich and R. M. Borisyuk, "Dynamics of neural networks with a central element," *Neural Networks*, 12, pp. 441-454, 1999.

[74] B. Khurana and E. Kowler, "Shared attentional control of smooth eye movement and perception," *Vision Research*, 27, pp. 1603-1618, 1987.

[75] C. Koch and S. Ullman, "Shifts in selective visual attention: towards the underlying neural circuitry," *Human Neurobiology*, 4, pp. 219-227, 1985.

[76] E. Kowler, J. van der Steen, EP. Tamminga and H. Collewijn, "Voluntary selection of the target for smooth eye movement in the presence of superimposed, full field stationary and moving stimuli," *Vision Search*, 24, pp. 1789-1798, 1984.

[77] E. Kowler and C. Iingale, "Smooth eye movements as indicators of selective attention," In M.I. Posner and O.S.M. M arin (Eds.) *Attention and Performance XI*, Hillsdale, NJ: Lawrence Erlbaum Associates Inc., pp. 285-300, 1985.

[78] E. Kowler, E. Anderson, B. Dosher and E. Blaser, "The role of attention in the programming of saccades," *Vision Research*, 35(13), pp. 1897-1916, 1995.

[79] A. F. Kramer and A. Jacobson, "Perceptual organization and focused attention: The role of objects and proximity in visual processing," *Perception and Psychophysics*, 50, pp. 267-284, 1991.

[80] V. I. Kryukov, "An attention model based on the principle of dominanta," In A. V. Holden and V. I. Kryukov (Eds.) *Neurocomputers and Attention I: Neurobiology, Synchronization and Chaos*, Manchester: Manchester University Press, pp. 319, 1991.

[81] H. W. Kwak, D. Dagenbach and H. Egeth, "Further evidence for a time-independent shift of the focus of attention", *Percept. Psychophys*, 49, pp. 473-80, 1991.

[82] D. LaBerge, *Attentional Processing: The Brain's Art of Mindfulness*, Harvard University Press, 1995.

[83] M. F. Land and S. Furneaux, "The knowledge base of the oculomotor system," *Philos. Trans. R. Soc. London Ser. B*, 352, pp. 1231-39, 1997.

[84] N. Lavie, "Perceptual load as a necessary condition for selective attention," *J. Exp. Psychol: Hum. Percept. Perf.*, 21, pp. 451-468, 1995.

[85] N. Lavie and J. Driver, "On the spatial extent of attention in object-based selection," *Perception and Psychophysics*, 58, pp. 1238-1251, 1996.

[86] F. L. Lim, G. A. W. West and S. Venkatesh, "Use of log polar space for foveation and feature recognition," *IEE Proceedings on Vision, Image and Signal Processing*, 144(6), pp. 323-331, December 1997.

[87] G. D. Logan, "The CODE theory of visual attention: an integration of space-based and object-based attention," *Psychological Review*, 103(4), pp. 603-649, 1996.

[88] S. J. Luck, "Neurophysiology of selective attention," In H. Pashler (Ed.), *Attention*, Psychology Press Ltd., pp. 257-295, 1998.

[89] A. Mack and I. Rock, *Inattentional Blindness*, Cambridge, MA: MIT Press, 1998.

[90] E. Ludvigh and J. W. Miller, "Study of visual acuity during the ocular pursuit of moving test objects: I. Introduction," *J. Opt. Soc. America*, 48, pp. 799-802, 1958.

[91] M. Mackeben and K. Nakayama, "Express attentional shifts", *Vision Research*, 33, pp. 85-90, 1993.

[92] A. Maki, P. Nordlund and J.-Q. Eklundh, "Attention scene segmentation: integrating depth and motion," *Computer Vision and Image Understanding*, 78, pp. 351-373, 2000.

[93] E. Matin, "Saccadic suppression: a review and an analysis," *Psychological Bulletin*, 81(12), pp. 899-917, 1974.

[94] R. M. McPeek, V. Maljkovic and K. Nakayama, "Saccades require focal attention and are facilitated by a short-term memory system," *Vision Research*, 39, pp. 1555-1566, 1999.

[95] R. Milanese, S. Gil and T. Pun, "Attentive mechanisms for dynamic and static scene analysis," Optical Engineering, 34(8), pp. 2428-34, 1995.

[96] K. Nakayama and M. Mackeben, "Sustained and transient components of focal visual attention," *Vision Research*, 29, pp. 1631-47, 1989.

[97] A. Nemcsics, *Color Dynamics*, Publisher Akadmiai Kiad, Budapest, 1993.

[98] U. Neisser, *Cognitive Psychology*, Englewood Cliffs, NJ: Prentice-Hall, 1967.

[99] E. Niebur, C. Koch and C. Rosin, "An oscillation based model for the neuronal basis of attention," *Vision Research*, 33, pp. 2789-2802, 1993.

[100] E. Niebur and C. Koch, "A model for the neuronal implementation of selective visual attention based on tempporal correlation among neurons," *J. Neurosci.*, 1, pp. 141-158, 1994.

[101] H. C. Nothdurft, "Feature analysis and the role of similarity in pre-attentive vision," *Perception and Psychophysics*, 52, pp. 355-375, 1992.

[102] H. C. Nothdurft, "The conspicuousness of orientation and motion contrast," *Spatial Vision*, 7(4), pp. 341-363, 1993.

[103] H. C. Nothdurft, "Focal attention in visual search," *Vision Research*, 39, pp. 2305-2310, 1999.

[104] B. A. Olshausen, C. H. Andersen and D. C. Van Essen, "A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information," *J. Neuroscience*, 13(11), pp. 4700-4719, 1993.

[105] S. E. Palmer, *Vision Science-Photons to Phenomenology*, Cambridge, MA: MIT Press, 1999.

[106] H. Pashler, *The Psychology of Attention*, Cambridge, MA: MIT Press, 1998.

[107] G. A. Patel and K. Sathian, "Visual search: bottom-up or top-down?" *Frontiers in Bioscience*, 5, d169-193, January 1, 2000.

[108] M. E. Posner, "Orienting of attention," *Q. J. Exp. Psychol.*, 32, pp. 3-25, 1980.

[109] M. E. Posner, Y. Cohen and R. D. Rafal, "Neural Systems control of spatial orienting," *Phil. Trans. R. Soc. Lond. B*, 298, pp. 187-198, 1982.

[110] M. I. Posner, J. A. Walker, F. J. Friedrich and R. D. Rafal, "Effects of parietal injury on covert orienting of attention", *Journal of Neuroscience*, 4, pp. 1863-1874, 1984.

[111] M. I. Posner and J. Driver, "The neurobiology of selective attention", *Curr. Opin. Neurobiol.*, 2, pp. 165-169, 1992.

[112] E. O. Postma, et al. "SCAN: A scalable model of attentional selection," *Neural Networks*, 10, pp. 993-1015, 1997.

[113] C. Poynton, "Frequently asked questions about color," http://www.inforamp.net/ Poynton/.

[114] Zenon W. Pylyshyn, *Seeing and Visualizing: It's Not What You Think (Life and Mind)*, Bradford Book, 2003.

[115] R. D. Rafal, "Balint syndrome," In T. Feinberg and M. Farah (Eds.), *Behavioral Neurology and Neuropsychology*, pp. 337-356, New York: McGraw-Hill, 1997.

[116] A. L. Ratan, "The role of fixation and visual attention in object recognition," MIT AI-TR-1529, July, 1995.

[117] G. Rizzolati, "Mechanisms of selective attention in mammals," In J. P. Ewart, R. R. Capranica and D. J. Ingle (Eds.), *Advances in vertebrate neuroethology*, Plenum, New York, pp. 261-297, 1983.

[118] A. S. Rojer and E. L. Schwartz, "Design Considerations for a Space-Variant Visual Sensor with Complex-Logarithmic Geometry," In *10th International Conference on Pattern Recognition*, 2, pp. 278-285, 1990.

[119] D. J. Robinson and S. E. Peterson, "The pulvinar and visual salience," *Trends in Neuroscience*, 15(4), pp. 127-132, 1992.

[120] I. A. Rybak, V. I. Gusakova, A. V. Golovan, L. N. Podladchikova and N. A. Shevtsova, "A model of attention-guided visual perception and recognition," *Vision Research*, 38, pp. 2387-2400, 1998.

[121] G. Sandini and M. Tristarelli, "Vision and Space-Variant Sensing," In H. Wechsler (Ed.), *Neural Networks for Perception*, Academic Press, pp. 398-425, 1992.

[122] B. J. Scholl, "Objects and attention: the state of the art," *Cognition,* 80, pp. 1-46, 2001.

[123] E. L. Schwartz, "Anatomical and physiological correlates of visual computation from striate to infero-temporal cortex," *IEEE Transactions on Systems, Man and Cybernetics*, SMC-14(2), pp. 257-271, 1984.

[124] C. Sears, *Inhibition of Return of Visual Attention and Visual Indexing*, PhD Dissertation, Department of Psychology, University of Western Ontario, 1995.

[125] G. Sela and M. D. Levine, "Real-time attention for robotic vision," *Real-Time Imaging*, 3, pp. 173-194, 1997.

[126] A. Shokoufandeh, I. Marsic and S. Dickinson, "View-based object recognition using saliency maps," *Image and Computing*, 17, pp. 445-460, 1999.

[127] A. M. Sillito, K. L. Grieve, H. E. Jones, J. Cudeiro and J. Davis, "Visual cortical mechanisms detecting focal orientation discontinuities," *Nature*, 378, pp. 492-496, 1995.

[128] W. Singer and C. W. Gray, "Visual feature integration and the temporal correlation hypothesis," *Annu. Rev. Neurosci.*, 18, pp. 555-86, 1995.

[129] F. Smeraldi and J. Bigun, "Retinal vision applied to facial feature detection and face authentication," *Pattern Recognition Letters*, 23, pp. 463-475, 2002.

[130] E. Spelke, W. Hirst and U. Neisser, "Skills of divided attention", *Cognition*, 4, pp. 215-230, 1976.

[131] G. Sperling and E. Weichselgartner, "Episodic theory of the dynamics of spatial attention," *Psychological Review*, 102, pp. 503-32, 1995.

[132] Y. Sun and R. Fisher, "Hierarchical Selectivity for Object-based Visual Attention," *Proc. 2nd Biologically Motivated Computer Vision Workshop (BMCV*

2002) Tuebingen, Germany, November 2002, pp. 427-438. Aka Springer LNCS 2525.

[133] Y. Sun and R. Fisher, "Object-based attention for computer vision," *Artificial Intelligence*, 146(1), pp. 77-123, 2003.

[134] Y. Sun, R. Fisher, F. Wang and H. M. Gomes, "Object-Based Attention-Driven Saccadic Eye Movements," submitted to *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2003.

[135] B. Takacs and H. Wechsler, "A dynamic and multiresolution model of visual attention and its application to facial landmark detection," *Computer Vision and Image Understanding*, 70(1), pp. 63-73, 1998.

[136] S. Tipper, B. Weaver, L. Jerreat and A. Burak, "Object-based and environment-based inhibition of return of visual attention," *Journal of Experimental Psychology: Human Perception and Performance*, 20, pp. 478-499, 1994.

[137] A. Treisman and G. Gelade, "A feature integration theory of attention," *Cognition Psychology*, 12, pp. 97-136, 1980.

[138] A. Treisman, "Perceptual grouping and attention in visual search for features and for objects," *J. Exp. Psychol: Hum. Percept. Perf.*, 8, pp. 194-214, 1982.

[139] A. Treisman, "Features and objects: the fourteenth Bartlett Memorial lecture," *Q. J. Experimential Psychology*, 40A, pp. 201-237, 1988.

[140] A. Treisman, "The perception of features and objects," In A. Baddeley and L. Weiskrantz (Eds.) *Attention: Selection, Awareness and Control*, Oxford: Uarendon Press, pp. 5-35, 1993.

[141] A. Treisman, "Feature binding, attention and object perception," *Phil. Trans. R. Soc. Lond. B.*, 353, pp. 1295-1306, 1998.

[142] J. K. Tsotsos, et al. "Modelling visual attention via selective tuning," *Artificial Intelligence*, 78, pp. 507-545, 1995.

[143] M. Usher and N. Donnelly, "Visual synchrony affects binding and segmentation in perception," *Nature*, 394, pp. 179-182, 1998.

[144] B. A. Wandell, "Computational neuroimaging: color representations and processing," In M. S. Gazzaniga (Ed.), *New Cognitive Neuroscience*, MIT Press, 1999.

[145] C. F. Westin, CJ. Westelius, H. Knutsson and G. Granlund, "Attention control for robot vision," *Proceedings of IEEE Computer Vision and Pattern Recognition*, San Francisco, CA, pp. 726-733, 1996.

[146] S. W. Wilson, "On the retino-cortical mapping," *Int. J. Man-Machine Studies*, 18, pp. 361-389, 1983.

[147] J. W. Wolfe, "Guided Search 2.0: A revised model of visual search," *Psychonomic Bulletin and Review*, 1, pp. 202-238, 1994.

[148] J. W. Wolfe, "Visual search," In H. Pashler (Ed.), *Attention*, Psychology Press Ltd., pp. 13-73, 1998.

[149] R. D. Wright and C. M. Richard, "Inhibition-of-return at multiple locations in visual space," *Canadian Journal of Experimental Psychology*, 50(3), pp. 324-327, 1996.

[150] Gunter Wyszechi, W. S. Stiles, Gunther Wyszecki and Gnther Wyszecki, *Color Science: Concepts and Methods, Quantitative Data and Formulae*, 2nd edition, published by John Wiley & Sons, 2000.

[151] S. Yantis, "Control of visual attention," In H. Pashler, (Ed.), *Attention*, Psychology Press Ltd., pp. 223-256, 1998.

[152] A. L. Yarbus, *Eye Movements and Vision*, New York: Plenum Press, 1967.

[153] R. S. Zemel, M. Behrmann, M. C. Mozer and D. Bevalier, "Experience-dependent perceptual grouping and object-based attention," *Journal of Experimental Psychology: Human Perception and Performance*, 28(1), pp. 202-217, 2002.